University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

1-1-2016

# Partial Information Framework: Basic Theory and Applications

Ville Antton Satopää
*University of Pennsylvania*, satopaa@wharton.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/edissertations

 Part of the Applied Mathematics Commons, Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation
Satopää, Ville Antton, "Partial Information Framework: Basic Theory and Applications" (2016). *Publicly Accessible Penn Dissertations*. 1991.
http://repository.upenn.edu/edissertations/1991

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/1991
For more information, please contact libraryrepository@pobox.upenn.edu.

# Partial Information Framework: Basic Theory and Applications

**Abstract**
Many real-world decisions depend on accurate predictions of some future outcome. In such cases the decision-maker often seeks to consult multiple people or/and models for their forecasts. These forecasts are then aggregated into a consensus that is inputted in the final decision-making process. Principled aggregation requires an understanding of why the forecasts are different. Historically, such forecast heterogeneity has been explained by measurement error. This dissertation, however, first shows that measurement error is not appropriate for modeling forecast heterogeneity and then introduces information diversity as a more appropriate yet fundamentally different alternative. Under information diversity differences in the forecasts stem purely from differences in the information that is used in the forecasts. This is made mathematically precise in a new modeling framework called the partial information framework. At its most general level, the partial information framework is a very reasonable model of multiple forecasts and hence offers an ideal platform for theoretical analysis. For one, it explains the empirical phenomenon known as extremization. This is a popular technique that often improves the out-of-sample performance of simple aggregators, such as the average or median, by transforming them directly away from the marginal mean of the outcome. Unfortunately, the general framework is too abstract for practical applications. To apply the framework in practice one needs to choose a parametric distribution for the forecasts and outcome. This dissertation motivates and chooses the multivariate Gaussian distribution. The result, known as the Gaussian partial information model, is a very close yet practical specification of the framework. The optimal aggregator under the Gaussian model is shown to outperform the state-of-the-art measurement error aggregators on both synthetic and many different types of real-world forecasts.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

**Graduate Group**
Statistics

**First Advisor**
Lyle H. Ungar

**Second Advisor**
Shane T. Jensen

**Keywords**
Expert belief, Forecast heterogeneity, Judgmental forecasting, Model averaging, Noise reduction, Probability modeling

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1991

**Subject Categories**
Applied Mathematics | Mathematics | Statistics and Probability

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1991

# PARTIAL INFORMATION FRAMEWORK: BASIC THEORY AND APPLICATIONS

Ville A. Satopää

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

**Supervisor of Dissertation**

Shane T. Jensen
Associate Professor of Statistics

**Co-Supervisor of Dissertation**

Lyle H. Ungar
Professor of Computer and Information Science

**Graduate Group Chairperson**

Eric Bradlow
K.P. Chao Professor, Marketing, Statistics and Education

**Dissertation Committee**

Shane T. Jensen, Associate Professor
Lyle H. Ungar, Professor
Edward I. George, Professor
Robin Pemantle, Professor
Philip E. Tetlock, Professor

PARTIAL INFORMATION FRAMEWORK: BASIC THEORY AND APPLICATIONS

COPYRIGHT © 2016

Ville A. Satopää

# Dedication

This dissertation has been dedicated to my mother, whose love and support give me courage to take on anything, and to my father, who lives on as the inspiration in everything I do.

Tämä tohtoruus on omistettu äidilleni, jonka rakkaus ja tuki antaa minulle rohkeutta kohdata ihan mitä vain, ja isälleni, joka elää inspiraationa kaikessa mitä teen.

# Acknowledgments

First and foremost, I would like to thank my family and friends whose unquestioning support has allowed me to chase my dreams to the edge of the world and back again. Second, I would like to thank my thesis committee for all their mentoring and advice about research, life, and everything in between. Lastly, I would like to thank my two Ph.D. advisors for always having my back, for reminding me of the bigger picture when I was lost in the details, and for making me a much better researcher. Thanks to them, I could not feel any better-equipped to begin my career as a professor.

Ensinnäkin haluan kiittää perhettäni ja ystäviäni joiden ehdoton tuki on antanut minulle mahdollisuuden jahdata haaveitani maailman reunalle asti ja takaisin. Toiseksi haluan kiittää tutkielmani komiteaa kaikesta mentoroinnista ja neuvoista liittyen tutkimustyöhön, elämään, ja kaikkeen siitä väliltä. Lopuksi haluan kiittää kahta tohtoruuden valvojaani siitä että ovat aina pitäneet puoltani, siitä että ovat muistuttaneet minua kokonaiskuvasta kun olen hävinnyt yksityiskohtiin, ja siitä että ovat tehneet minusta paljon paremman tutkijan. Heidän ansiosta en voisi tuntea itseäni enemmän valmiiksi aloittamaan urani professorina.

# ABSTRACT

# PARTIAL INFORMATION FRAMEWORK: BASIC THEORY AND APPLICATIONS

Ville A. Satopää

Shane T. Jensen

Lyle H. Ungar

Many real-world decisions depend on accurate predictions of some future outcome. In such cases the decision-maker often seeks to consult multiple people or/and models for their forecasts. These forecasts are then aggregated into a consensus that is inputted in the final decision-making process. Principled aggregation requires an understanding of why the forecasts are different. Historically, such forecast heterogeneity has been explained by measurement error. This dissertation, however, first shows that measurement error is not appropriate for modeling forecast heterogeneity and then introduces information diversity as a more appropriate yet fundamentally different alternative. Under information diversity differences in the forecasts stem purely from differences in the information that is used in the forecasts. This is made mathematically precise in a new modeling framework called the partial information framework. At its most general level, the partial information framework is a very reasonable model of multiple forecasts and hence offers an ideal platform for theoretical analysis. For one, it explains the empirical phenomenon known as extremization. This is a popular technique that often improves the out-of-sample performance of simple aggregators, such as the average or median, by transforming them directly away from the marginal mean of the outcome. Unfortunately, the general framework is too abstract for practical applications. To apply the framework in practice one needs to choose a parametric distribution for the forecasts and outcome. This dissertation motivates and chooses the multivariate Gaussian distribution. The result, known as the Gaussian partial information model, is a very close yet practical specification of the framework. The optimal aggregator under the Gaussian model is shown to outperform the state-of-the-art measurement error aggregators on both synthetic and many different types of real-world forecasts.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

A new form of polling has emerged from the recent development of computer and social networks; it is called prediction polling (Atanasov et al., 2015). In a prediction poll a group of participants collectively make predictions about some future quantity of interest. For instance, consider a policy maker who is interested in the probability of Brexit. Instead of collecting data and aiming to build a statistical model, the policy-maker may reach out to a group of European Union experts and ask them for their subjective probabilities of the event. After this, the decision-maker must choose how to use the forecasts. The first idea may be to simply follow the most accurate or informed forecaster's advice. Unfortunately, however, it is often not possible to know ex-ante who this forecaster is, and even if one somehow could know, simply following a single forecaster's advice would ignore potentially a large amount of information that is being contributed by the rest of the forecasters. Therefore a better option is to combine the forecasts into a single consensus forecast that reflects all forecasters' information. Unfortunately, there are many ways one could combine the predictions, and the final combination rule will largely determine the out-of-sample performance of the consensus forecast.

The past literature has distinguished two general approaches to forecast aggregation:

1. **Empirical.** Overall, this approach is by far the more widely studied one. It is akin to machine learning in a sense that the decision-maker first picks some class of aggre-

gators and within that class chooses the aggregator that performs the best over some training set of past predictions on known outcomes. The chosen aggregator is then used for combining any future predictions of unknown outcomes.

2. **Model-based.** This approach begins with a probability model of forecast heterogeneity, that is, the way the predictions differ from each other and the outcome. The model-based aggregator is then the optimal aggregator under this assumed outcome-forecast link. Note that applying the aggregator in practice may or may not involve estimating some model parameters from the forecasts – but not from the outcomes.

Both of these approaches are important and serve somewhat different purposes. More specifically, the empirical techniques are often simpler and work very well when one has access to a training set that is representative of the future aggregation tasks. There are, however, many forecasting applications where a training set is not available. For instance, in prediction polling obtaining a training set would require a lot of time and effort on behalf of the forecasters and polling agency. For this reason, many prediction polls do not yield a training set. Instead, the participants are typically handed out a single questionnaire that solicits their predictions about one or more future outcomes. Fortunately, the model-based aggregation approach can be applied directly to the forecasts even when no knowledge of the outcomes is available. Therefore the model-based approach is much more broadly applicable than the empirical approach. Furthermore, the model-based aggregators are based on theory which provides a clear direction for improvement. Of course, this all comes at a cost. In particular, the model-based approach relies on modeling assumptions. If these assumptions are no appropriate, the resulting aggregators are, unfortunately, only of limited use. Therefore it is important to perform careful model evaluation of any proposed model-based aggregators.

This dissertation was largely motivated by the lack of an appropriate framework for model-based aggregation. Historically, possibly due to early forms of data collection, fore-

cast heterogeneity has been explained with measurement error: the forecasts are assumed to be equal to the true outcome plus some mean zero idiosyncratic error. While this assumption may make sense when modeling estimates arising from repeated applications of a sensitive yet somewhat imprecise instrument, it is hardly reasonable when the estimates arise from multiple, often widely different sources. Furthermore, the measurement error based aggregators are different types of measures of central tendency such as the (weighted) average or median. Unfortunately, such simple aggregators do not behave as if they are collecting information from the different forecasters. To illustrate, consider a patient who is worried about his or her health and hence goes to the hospital. At the hospital both a blood test and an MRI are taken. Both tests come back with no evidence of poor health. Suppose there are two doctors who decide to look at the patient's case: doctors $A$ and $B$. Doctor $A$ only looks at the blood test results and provides a probability of $0.9$ of the patient being healthy. Doctor $B$, on the other hand, only looks at the MRI results but also provides a probability of $0.9$ of the patient being healthy. Now, the patient has two $0.9$s that are based on very different information. How should they be aggregated? Surely, if one were to see the good news both from the blood test and the MRI, one would be even more convinced of the patient's good health and hence predict something greater than $0.9$. In other words, in this simple example the combined evidence should yield a forecast somewhat greater than $0.9$. Unfortunately, however, all measures of central tendency aggregate precisely to $0.9$. Therefore they fail to account for the doctors' differing information sets and hence cannot collect information from the different forecasters. Of course, this is only a simple example. The result, however, is much more general than this. In fact, as will be shown in Chapter 7, this result holds for any number of forecasters despite whether their forecasts are equal or not.

In order to fix this shortcoming, it is necessary to revisit the fundamentals. In particular, a new source of forecast heterogeneity must be introduced. This should be more appropri-

ate than measurement error and it should lead to aggregators that do behave as if they were collecting information from the different forecasters. Such an alternative is precisely what is introduced in this dissertation; it is called *information diversity*. Under information diversity, the differences in the predictions are fully explained by differences in the information used by the respective forecasters. For instance, consider two forecasters predicting the chances of some global crisis. One of these forecasters lives in USA and follows the American news. The second forecaster, on the other hand, lives in Russia and reads the Russian news. Given these descriptions, it is likely that the two forecasters have different information and hence provide different predictions. This intuition is mathematically formalized in a new modeling framework called the *partial information framework*. Overall, the partial information framework is very general and can be applied to a broad range of different forecasting applications, allowing the practitioner to construct application-specific aggregators instead of always relying on the usual average and median. The partial information aggregators also behave as if they collect information from the forecasters and often outperform the state-of-the-art measurement error aggregators in real-world applications.

Given that information diversity is a contribution at the root of statistical theory, it gives rise to a large amount of new theory and methodology. This dissertation discusses several such projects. In particular, each chapter is a separate paper discussing a different aspect of the partial information framework. The chapters have been ordered chronologically in the order they were written. The following enumeration briefly describes each chapter and provides citations of the corresponding papers.

**Chapter 2.** Satopää et al. (2014) introduces a new empirical aggregator for probability forecasts. Overall, the aggregator is very simply: it involves only a single tuning parameter which determines how much the average log-odds forecast should be extremized. Here extremization refers to the process of transforming a measure of central tendency, such as the average, directly away from the marginal mean (typi-

cally at $0.5$ for probability forecasts) and closer to the nearest extreme (at $0.0$ or $1.0$). This extremizing aggregator is then shown to outperform simple measurement-error aggregators on real-world predictions. Furthermore, the amount of extremization was observed to decrease in the forecaster's self-reported expertise.

*Note:* At this point the benefits of extremizing were merely an empirical observation. In particular, it was not clear why it helps or how much extremization should be performed.

**Chapter 3.** Satopää et al. (2014)[1] was largely motivated by the data collected by the Good Judgment Project (GJP) (Mellers et al., 2014). More specifically, the GJP recruited 1,000s of experts to make probability forecasts of hundreds of future events deemed important by the Intelligence Advanced Research Projects Activity (IARPA). Each event was succeeded by a period of time during which the forecasters were allowed to make predictions and update them if they felt that the likelihoods had changed. Therefore each forecaster gave a time-series of predictions. To aggregate such streams of forecasts, this paper develops an empirical aggregator that extremizes and combines predictions over time. Furthermore, the amount of extremization is allowed to vary across different self-reported expertise groups. Overall, the aggregator outperforms classical exponentially-weighted aggregators on real-world predictions from the GJP.

*Note:* Even though this paper contributes to a common forecasting setup where the forecasters are allowed to update their forecasts over time, all methodology is empirical and hence requires the decision-maker to conduct a rather large study in which the forecasters are making and updating forecasts for multiple events. Also, given that the aggregator is learned over a training set, the decision-maker must wait for

[1]This is the thesis for my Master of Arts in Statistics degree received in December 2014. It is included here for completeness.

each of the events to be resolved. This illustrates the limitations of the empirical aggregation approach. In some sense, the benefits of extremizing suggest a bias in the underlying probability model, namely the measurement error model that motivates the simple aggregators applied before extremization.

**Chapter 4.** Satopää et al. (2015) introduces the partial information framework for probability forecasts. The framework motivates two benchmarks for aggregation: the oracular and revealed aggregators. The oracular aggregator has access to all the details of the forecasters' information and hence is only useful for theoretical analysis. The revealed aggregator, on the other hand, can be used in practice as it only depends on information revealed through the reported forecasts. As a practical specification of the framework the first version of the Gaussian partial information model is developed. This version describes full information by some closed interval. Each forecaster then observes some Borel subset of this interval. The variance of the forecast is the size of the corresponding Borel set, and the covariance between any two forecasters is the size of the overlap between their Borel sets. This motivates a structure for the covariance matrix that has to be in a convex set known as the correlation polytope. In the paper the oracular aggregator under the Gaussian model is used as benchmark to analyze the amount of required extremization under different information structures. In particular, it is found that extremization increases in two separate measures: the total amount of information and the amount of information diversity among the forecasters. This motivates a spectrum of aggregators, ranging from averaging (full information overlap) to summing (no information overlap).

*Note:* While information diversity is intuitively much more appealing than measurement error, this paper was mainly theoretical and contained very little empirical evidence in favor of information diversity. Furthermore, the focus was entirely on modeling probability forecasts. Based on the general analysis, however, it is clear that

information diversity is a much more general concept: it is an alternative to measurement error and hence suggests a more general modeling framework.

**Chapter 5.** Satopää et al. (2016) introduces information diversity as a general alternative to measurement error and shows how the partial information framework can be used in practice to model different types of outcome-forecast pairs, such as probability forecasts of binary outcomes or real-valued forecasts of real-valued outcomes. Such applications are made more tractable by modifying the Gaussian model. In particular, it turns out that applying the partial information framework in practice only needs a choice of a parametric family of distributions for the outcome and forecasts – nothing else. The multivariate Gaussian distribution here leads to our second version of the Gaussian model. This time the form of the forecasts' covariance matrix is motivated solely by the general partial information framework instead of overlapping Borel sets. Most importantly, however, the revised form is much more tractable than the one introduced by the first version of the Gaussian model. The paper then develops a procedure for estimating these covariance matrices and applies the revised Gaussian model to two real-world applications. The analysis leads to several observations. First, in both cases the revealed aggregator significantly outperforms the state-of-the-art measurement error aggregators. Second, unlike the measurement error aggregators, the revealed aggregator behaves as if it is collecting information from the forecasters. Third, the estimated information structure aligns well with prior knowledge about the forecasters' information.

*Note:* This paper provides much empirical evidence in favor of information diversity as the more important source of forecast heterogeneity. The estimation procedure therein, however, requires the forecasters to make predictions for multiple related outcomes. Unfortunately, in many forecasting setups multiple outcomes are not available, and even if they are, the outcomes may different in nature such that the

information structure cannot be assumed to remain constant among them. Ideally, one would have a partial information aggregator that can operate directly on a set of forecasts of a single outcome.

**Chapter 6.** Ernst et al. (2016) introduces the notion of Bayes-Gaussian aggregation. The motivation relies on previous applications that have inputted a point estimate of the information structure to the revealed aggregator. Such plug-in aggregators, however, can be unstable. To avoid this, a Bayes-Gaussian aggregator computes a posterior-weighted mixture of the plug-in aggregators under all possible information structures. In this paper a Bayes-Gaussian aggregator is developed for two probability forecasts of a single event. To simplify the computations, each of the forecasters is assumed to know half of the total information. Their information overlap, however, is considered unknown and is analytically integrated out with respect to its posterior distribution. The final form is a simple aggregator that is free of any model parameters and hence can be applied directly to any two probability forecasts. Even though the literature on Bayesian statistics offers many numerical procedures for integration, the model parameters here are integrated out analytically in order to arrive at a closed-form aggregator. The hope is that such a simple closed-form encourages practitioners to abandon the usual average and median aggregators.

**Chapter 7.** Satopää and Ungar (2015) begins by proving and discussing some general results under the partial information framework. For one, it makes the notion of information collection precise: an information collector is a calibrated aggregator, as this shows that it is consistent with some information set about the outcome, and its variance is at least as large as the maximum variance among the individual forecasts. The paper then proceeds to show that the revealed aggregator does collect information. In contrast, the weighted average of calibrated forecasts is shown to be neither calibrated nor to collect information. Given that measures of central

tendency generally reduce variance, these simple aggregators can be intuitively seen to not collect information either. Furthermore, the weighted average tends to be too close (as compared to the revealed aggregator) to the marginal mean. This motivates extremization of the weighted average under any type of forecast-outcome pair. The paper concludes by showing how extremizing real-valued forecasts improves both calibration and prediction accuracy under synthetic and real-world data.

# 2

# Combining Multiple Probability Predictions Using a Simple Logit Model*


## Abstract

This paper first presents a simple model of how experts estimate probabilities. The model is then used to construct a likelihood-based aggregation formula for combining multiple probability forecasts. The resulting aggregator has a simple analytic form that depends on a single easily-interpretable parameter. This makes it computationally simple, attractive for further development, and robust against overfitting. Based on a large-scale dataset in which over 1,300 experts tried to predict 69 geopolitical events, our aggregator is found to be superior to several widely used aggregation algorithms.


## 2.1 Introduction

Experts are often asked to give decision makers subjective probability estimates on whether certain events will occur or not. After collecting such probability forecasts, the challenge is to construct an aggregation method that produces a consensus probability for each event

*Joint work with Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, Lyle H. Ungar

by combining the probability estimates appropriately. If the observed long-run empirical distribution of the events matches the aggregate forecasts, the aggregation method is said to be calibrated. This means that, for instance, 30% of the events, which have been assigned a aggregate forecast of 0.3, occur. According to Ranjan (2009), however, calibration is not sufficient for useful decision making. The aggregation method should also maximize *sharpness* which increases as the aggregate forecasts concentrate closer around the extreme probabilities 0.0 and 1.0. Therefore it can be said that the overall goal in probability estimation is to maximize sharpness subject to calibration (for more information see, e.g., Gneiting et al. 2007; Pal 2009).

The most popular choice for aggregation is *linear opinion pooling*, which assigns each individual forecast a weight reflecting the importance of the expert. Ranjan and Gneiting (2010), however, show that any linear combination of (calibrated) forecasts is uncalibrated and lacks sharpness. Furthermore, Allard et al. (2012) show in several simulations studies that linear opinion pooling performs poorly relative to other pooling formulas with a multiplicative instead of an additive structure.

Previous literature has introduced a wide range of methods that aggregate probabilities in a non-linear manner (see, e.g., Ranjan and Gneiting 2010; Bordley 1982; Polyakova and Journel 2007). Many of these methods, however, involve a large number of parameters making them computationally complex and susceptible to over-fitting. By contrast, parameter-free approaches such as the median or the geometric mean of the odds are too simple to optimally incorporate the use of training data. In this paper, we propose a novel aggregation approach that is simple enough to avoid over-fitting, straightforward to implement, and yet flexible enough to make use of training data. Therefore our aggregator retains the benefits of parsimony from parameter-free approaches without losing the ability to use training data.

The theoretical justification for our aggregator arises from a log-odds statistical model

of the data. The log-odds representation is convenient from a modeling perspective. Being defined on the entire real line, the log-odds can be modeled with a Normal distribution. For example, Erev et al. (1994) model log-odds with a Normal distribution centered at the "true log-odds"[2]. The variability around the "true log-odds" is assumed to arise from the personal degree of momentary confidence that affects the process of reporting an overt forecast. We extend this approach by adding a *systematic bias* component to the Normal distribution. That is, the Normal distribution is centered at the "true log-odds" that have been multiplied by a small positive constant (strictly between zero and one) and are hence systematically regressed toward zero.

To illustrate this choice of location, assume that 0.9 is the most informed probability forecast that could be given for a future event with two possible outcomes. A rational forecaster who aims to minimize a reasonable loss function, such as the Brier score[3], without any previous knowledge of the event, gives 0.5 as his initial probability forecast. However, as soon as the forecaster gains some knowledge about the event, he produces an updated forecast that is a compromise between his initial forecast and the new information acquired. The updated forecast is therefore conservative and necessarily too close to 0.5 as long as the forecaster remains only partially informed about the event. If most forecasters fall somewhere on this spectrum between ignorance and full information, their average forecast tends to fall strictly between 0.5 and 0.9 (see Baron et al. (2014) for more details). This discrepancy between the "true probability" and the average forecast is represented in our model by using the regressed "true log-odds" as the center of the Normal distribution.

Both Wallsten et al. (1997) and Zhang and Maloney (2012) recognize the presence of this systematic bias. Wallsten et al. (1997) discuss a model with a bias term that regresses

[2]In this paper, we use quotation marks in any reference to a true probability (or log-odds) to avoid a philosophical discussion. These quantities should be viewed simply as model parameters that are subject to estimation.

[3]The Brier score is the squared distance between the probability forecast and the event indicator that equals 1.0 or 0.0 depending on whether the event happened or not, respectively.

the expected responses towards 0.5. Zhang and Maloney (2012) provide multiple case studies showing evidence for the existence of the bias. Neither study, however, describe a way of correcting the bias or a potential aggregation method to accompany the correction. Zhang and Maloney (2012) estimate the bias at an individual level requiring multiple probability estimates from a single forecaster. Even though our approach can be extended rather trivially to correct the bias at any level (individual, group, or collective), in this paper we treat the experts as being indistinguishable and correct the systematic bias at a collective level by shifting each probability forecast closer to its nearest boundary point. That is, if the probability forecast is less (or more) than 0.5, it is moved away from its original point and closer to 0.0 (or 1.0).

This paper begins with the modeling assumptions that form the basis for the derivation of our aggregator. After describing the aggregator in its simplest form, the paper presents two extensions: the first one generalizes the aggregator to events with more than two possible outcomes, and the second one allows for varying levels of systematic bias at different levels of expertise. The aggregator is then evaluated under multiple synthetic data scenarios and on a large real-world dataset. The real data were collected by recruiting over 1,300 forecasters ranging from graduate students to forecasting and political science faculty and practitioners, and then posing them 69 geopolitical prediction problems (see the Appendix for a complete listing of the problems and Ungar et al. 2012 for more details on the data collection process). Our main contribution arises from our ability to evaluate competing aggregators on the largest dataset ever collected on geopolitical probability forecasts made by human experts. With such a large dataset, we have been able to develop a generic aggregator that is analytically simple and yet outperforms other widely used competing aggregators in practice. After presenting the evaluation results, the paper concludes by exploring some future research ideas.

## 2.2 Theory

Using the logit function

$$\text{logit}(p) \quad = \quad \log\left(\frac{p}{1-p}\right)$$

a probability forecast $p \in [0,1]$ can be uniquely mapped to a real number called the log-odds, $\text{logit}(p) \in \mathbb{R}$. This allows us to conveniently model probabilities with well-studied distributions, such as the Normal distribution, that are defined on the entire real line. In this section, assume that we have $N$ experts each giving *one* probability forecast for a binary-outcome event. We consider these experts as interchangeable. That is, no forecaster can be distinguished from the others either across or within problems. Denote the experts' forecasts with $p_i$ and let $Y_i = \text{logit}(p_i)$ for $i = 1, 2, \ldots, N$. As discussed earlier, we model the log-odds with a Normal distribution centered at the "true log-odds" that have been regressed towards zero by a factor of $a$. More specifically,

$$Y_i \quad = \quad \log\left(\frac{p}{1-p}\right)^{1/a} + \epsilon_i,$$

where $a \geq 1$ is an unknown level of systematic bias, $p$ is the "true probability" to be estimated, and each $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ is a random shock with unknown variance $\sigma^2$ on the individual's reported log-odds. If the model is correct, the event arising from this model would occur with probability $p$. Therefore $p$ should be viewed as a model parameter that is subject to estimation.

The larger $a$ is, the more the log-odds are regressed towards $0$ or, equivalently, the more the probability estimates are regressed towards $0.5$. Therefore we associate $a = 1$ with an accurate forecast and any $a > 1$ with a partially informed and under-confident forecast (Baron et al., 2014). It is certainly possible for an expert to be overconfident (see, e.g., McKenzie et al. 2008 for a recent and comprehensive discussion). In fact, we find this to

be the case among forecasters at the highest level of self-reported expertise. In Section 2.3.3.3 we provide empirical evidence that the forecasters as a group, however, tend to be under-confident. We therefore treat group-level under-confidence as a reasonable modeling restriction that we do not need to impose in our simulations (see Section 2.3), where we allow the data to speak for themselves by letting $a \in [0, \infty)$.

Notice that, unlike the systematic bias term $a$, the random error component $\epsilon_i$ is allowed to vary among experts. Putting this all together gives

$$\begin{aligned}
& \log\left(\frac{p_i}{1-p_i}\right) \overset{i.i.d.}{\sim} \text{Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right) \\
\Leftrightarrow \quad & \frac{p_i}{1-p_i} \overset{i.i.d.}{\sim} \text{Log-Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right) \\
\Leftrightarrow \quad & p_i \overset{i.i.d.}{\sim} \text{Logit-Normal}\left(\log\left(\frac{p}{1-p}\right)^{1/a}, \sigma^2\right)
\end{aligned}$$

This model is clearly based on an idealization of the real world and is therefore an over-simplification. Although performing a formal statistical test to determine whether the log-odds in our real-world dataset follow a Normal distribution lead to rejection of the null hypothesis of normality, this result simply reflects the inevitability of slight deviation from normality and the sensitivity of the statistical tests involving large sample sizes. Assuming normality, however, turns out to be a good enough approximation to be of practical use. While Zhang and Maloney (2012) did not model log-odds with a Normal distribution, they argue in favor of using the logit-transformation with a linear bias term to model probabilities. Di Bacco et al. (2003) use the Logit-Normal distribution to jointly model experts' probabilities under different levels of information. For our purposes, the Logit-Normal model serves as a theoretical basis for a clean and justified construction of an efficient aggregator.

### 2.2.1 Model-based Aggregator

The invariance property of the maximum likelihood estimator (MLE) can be used to show that the MLE of $p$ is

$$\hat{p}_G(a) \;=\; \frac{\exp\left(a\bar{Y}\right)}{1+\exp\left(a\bar{Y}\right)},$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. By plugging in the definition of $Y_i$, the MLE can be expressed in terms of the geometric mean of the odds as

$$\hat{p}_G(a) \;=\; \frac{\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{1/N}\right]^a}{1+\left[\prod_{i=1}^{N}\left(\frac{p_i}{1-p_i}\right)^{1/N}\right]^a}, \tag{2.1}$$

where the subindex $G$ indicates the use of the geometric mean. The input argument emphasizes the dependency on the unknown quantity $a$. The estimator $\hat{p}_G$ is particularly convenient because it allows for (i) an easy extension to uneven expert weights by simply replacing each $1/N$ with a weight term $w_i$ and (ii) switching the order of transformation and aggregation operators. Notice, however, that making use of (i) would result in an estimator with a total of $N$ parameters. Such an estimator would be computationally complex and susceptible to overfitting. Many authors including Graefe et al. (2014a), Armstrong (2001), and Clemen (1989) encourage the use of equal weights unless there is strong evidence to support unequal weightings of the experts. For simplicity, we limit this paper to the equally weighted aggregator.

### 2.2.2 Estimating Systematic Bias

Our aggregator $\hat{p}_G$ depends on the unknown quantity $a$, which needs to be inferred. If we have a training set consisting of $K$ binary-outcome events and $K$ pools of probability

forecasts associated with these events, we can measure the goodness of fit for any $a$ with the mean score

$$\bar{S}_K(a) = \frac{1}{K}\sum_{k=1}^{K} S(\hat{p}_{G,k}(a), Z_k),$$

where $S$ is a proper scoring rule (see, e.g., Gneiting and Raftery (2007)), $\hat{p}_{G,k}$ is the aggregate probability forecast for the $k$th event, and the event indicator $Z_k \in \{0, 1\}$ depending on whether the $k$th event occurred ($Z_k = 1$) or did not occur ($Z_k = 0$). Optimizing this mean score as a function of $a$ gives the *optimum score estimator*

$$\hat{a}_{OSE} \quad = \quad \arg\min_a \bar{S}_K(a),$$

which according to Gneiting and Raftery (2007) is a consistent estimator of $a$.

Although strictly proper scoring rules are the natural loss functions in estimating binary class probabilities (see Buja et al. (2005)), the real appeal arises from the freedom of choosing a proper scoring rule to suit the problem at hand. Among the infinite number of proper scoring rules, the two most popular ones are the Brier score (see Brier 1950) and the logarithmic scoring rule (see Good 1952), which is equivalent to maximizing the log-likelihood and hence finding the maximum likelihood estimator of $a$. Given that it is not clear which rule should be used for predicting social science events, we estimate $a$ both via the Brier score

$$\hat{a}_{BRI} \quad = \quad \arg\min_a \sum_{k=1}^{K} \left(\hat{p}_{G,k}(a) - Z_k\right)^2$$

and via the likelihood function

$$\hat{a}_{MLE} \quad = \quad \arg\max_a \prod_{k=1}^{K} \hat{p}_{G,k}(a)^{Z_k}(1 - \hat{p}_{G,k}(a))^{1-Z_k},$$

and compare the resulting two aggregators. Notice that both equations are non-linear optimization problems with no analytic solutions. Fortunately, the optimizing values can be found with numerical methods such as the Newton-Raphson method or a simple line search.

### 2.2.3 Extensions to the Aggregator

This section briefly discusses two extensions to the aggregator. The first one extends $\hat{p}_G$ to events with more than two possible outcomes. This gives a more general aggregator with $\hat{p}_G$ as a sub-case. The second extension allows for varying values of $a$ across different groups of expertise.

#### 2.2.3.1 Multinomial Events

For now, assume that the event can take exactly one of a total of $M \geq 2$ different outcomes. Under pure ignorance the forecaster should assign $1/M$ probability to each outcome. The more ignorant the forecaster is, the more we would expect him to shrink his forecasts towards $1/M$. See Zhang and Maloney (2012) and Fox and Rottenstreich (2003) for further discussion.

We use this idea to generalize our aggregator. Choosing the $M$th outcome as the baseline, denoting the forecast for the $m$th outcome by the $i$th forecaster with $p_{i,m}$, and letting $Y_{i,m} = \log\left(\frac{p_{i,m}}{p_{i,M}}\right)$ for $i = 1, 2, \ldots, N$, we arrive at a more general version of the model represented as

$$Y_{i,m} \;=\; \log\left(\frac{p_m}{p_M}\right)^{1/a} + \epsilon_{i,m}$$

where $m \in \{1, \ldots, M-1\}$ and $\epsilon_{i,m} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma^2$ unknown. The resulting

maximum likelihood estimator for the $k$th outcome is

$$\hat{p}_{G,k}(a) = \frac{\left[\prod_{i=1}^{N}\left(\frac{p_{i,k}}{p_{i,M}}\right)^{1/N}\right]^{a}}{\sum_{j=1}^{M}\left[\prod_{i=1}^{N}\left(\frac{p_{i,j}}{p_{i,M}}\right)^{1/N}\right]^{a}}$$

Instead of analyzing this more general estimator, this paper will focus on the binary subcase. Notice, however, that all the properties generalize trivially to the multi-outcome case.

#### 2.2.3.2 Correction under Levels of Expertise

The reasoning in the previous subsection suggests that better forecast performance could be achieved by correcting for systematic bias differently at different levels of expertise. To make this more specific, assume that each forecaster can identify himself with one of $C$ expertise levels with $C$ being the most knowledgeable. Let $\boldsymbol{a} = [a_1, \ldots, a_C]'$ be a vector of $C$ different systematic bias factors, one for each expertise level. Then,

$$\hat{p}_{G,k}(\boldsymbol{a}) = \frac{\prod_{i=1}^{N}\left(\frac{p_{i,k}}{p_{i,M}}\right)^{\frac{\boldsymbol{e}_i'\boldsymbol{a}}{N}}}{\sum_{j=1}^{M}\prod_{i=1}^{N}\left(\frac{p_{i,j}}{p_{i,M}}\right)^{\frac{\boldsymbol{e}_i'\boldsymbol{a}}{N}}}$$

where $\boldsymbol{e}_i$ is a vector of length $C$ indicating which level of expertise the $i$th forecaster belongs to. For instance, if $\boldsymbol{e}_i = [0, 1, 0, \ldots, 0, 0]'$, the $i$th expert identifies himself with expertise level two. The systematic bias factors can be estimated by first partitioning the dataset by expertise and then finding the optimal value for each expert group separately. We will return to this topic briefly at the end of the results section, where we show the effects of actual expertise self-ratings from forecasters.

## 2.3 Results and Discussion

This section compares different aggregators both on synthetic and real-world data. The aggregators included in the analysis are as follows.

(a) Arithmetic mean of the probabilities

(b) Median of the probabilities

(c) Logarithmic opinion pool

$$\hat{p} \;=\; \prod_{i=1}^{N} p_i^{w_i} \Bigg/ \left( \prod_{i=1}^{N} p_i^{w_i} + \prod_{i=1}^{N} (1-p_i)^{w_i} \right),$$

which according to Bacharach (1972) was proposed by Peter Hammond (see Genest and Zidek (1986)). Given that we consider the forecasters indistinguishable, we assign equal weights to each forecaster. Letting $w_i = 1/N$ for $i = 1, \ldots, N$ gives us the equally weighted logarithmic opinion pool (ELOP).

(d) Our aggregator $\hat{p}_G(a)$ as given by Equation (2.1)

(e) The Beta-transformed linear opinion pool

$$\hat{p}(\alpha, \beta) \;=\; H_{\alpha,\beta}\left( \sum_{i=1}^{N} w_i p_i \right),$$

where $H_{\alpha,\beta}$ is the cumulative distribution function of the Beta distribution with parameters $\alpha$ and $\beta$, and $w_i$ is the weight given to the $i$th forecast. Allard et al. (2012) show, on simulations, that Beta-transformed linear pooling presents very good forecast performance. Again we assign equal weights to each forecaster. Letting $w_i = 1/N$ for $i = 1, \ldots, N$ gives us the Beta-transformed equally weighted linear opinion pool (BELP). This aggregator, however, tends to overfit in all of our evaluation procedures. Much better performance is obtained by requiring $\alpha = \beta \geq 1$. Un-

der such a restriction, the BELP aggregator can be enforced to shift any mean probability more toward the closest extreme probability 0.0 or 1.0. This one-parameter sub-case (1P-BELP) is more robust against overfitting and is supported by the theoretical results by Wallsten and Diederich (2001). For these reasons, it is a good competing aggregator in our simulations. We do not present the results associated with the 2-parameter BELP aggregator because BELP performs much worse than 1P-BELP in all of our simulations.

As suggested by Ranjan and Gneiting (2010), the parameter $\alpha$ can be fit by using optimum score techniques. We fit any tuning parameters using both the Brier score and the likelihood function, and then compare the resulting aggregators. Given that Ranjan and Gneiting (2010) only considered aggregating binary events, it is unclear how the Beta-transformed linear pooling can be generalized to events with more than two possible outcomes. Therefore our comparison will focus only on forecasting binary events.

Throughout this evaluation section, we will be using the Brier score as the performance measure. As discussed earlier in Section 2.2.2, this scoring rule has some attractive properties and is in essence a quadratic penalty. It also has an interesting psychological interpretation as the expected cost of an error, given a probability judgment and the truth (see Baron et al. 2014 for the details).

### 2.3.1 Synthetic Data: Correctly Specified Model

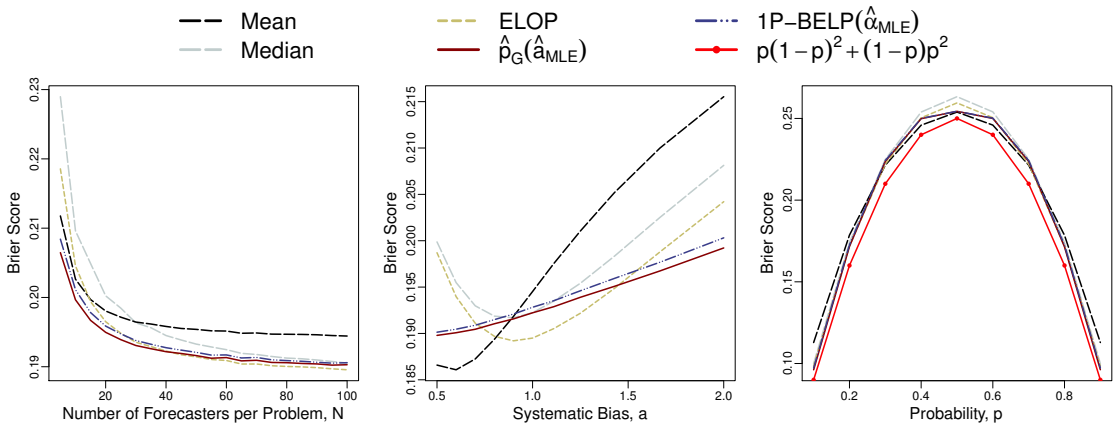
In this section we evaluate the different aggregators on a correctly specified model; that is, on data that have been generated directly from the Logit-Normal distribution described in Section 3.3. The evaluation is done over a three-dimensional grid that expands the number of forecasters per problem, $N$, from 5 to 100 (with increments of five forecasters), the true probability, $p$, from 0.1 to 0.9 (with increments of 0.1), and the systematic bias term, $a$, from $5/10, 6/10, \ldots 9/10, 10/10, 10/9, \ldots, 10/6, 10/5$ symmetrically around the no-bias

point at 1.0. The simulation was run for 100 iterations at every grid point. Each iteration used the values at the grid point to produce a synthetic data set from the Logit-Normal distribution. The true probability, $p$, was used to generate Bernoulli random variables that indicated which events occurred and which did not. Testing was performed on a separate testing set consisting of 1,000 problems, each with the same number of forecasters as the problems in the original training set. The simulation was repeated for two different numbers of problems in the training set, $K = 30$ and $K = 100$. The variance for the log-odds, $\sigma^2$, was equal to $5$ throughout the entire simulation[4].

The results are summarized in two sets of figures: Figures 2.1a and 2.2a plot the Brier scores (given by averaging over the systematic bias and the true probability) against the number of forecasters per problem, Figures 2.1b and 2.2b plot the Brier score (given by averaging over the number of forecasters per problem and the true probability) against the systematic bias term, and Figures 2.1c and 2.2c plot the Brier scores (given by averaging over the number of forecasters per problem and the systematic bias) against the true probability. Figure 2.1 presents the results under $K = 30$, and Figure 2.2 shows the results under $K = 100$.

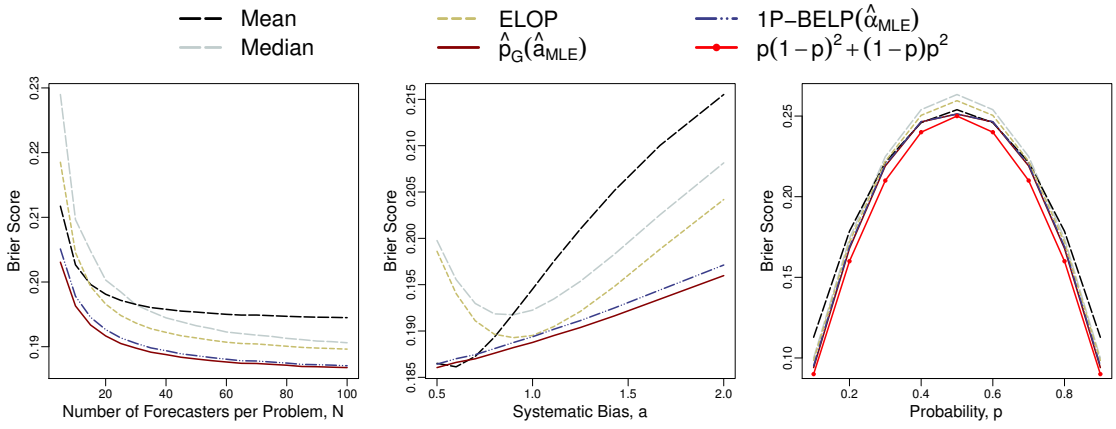
Given that $\hat{p}_G(\hat{a}_{MLE})$ and 1P-BELP$(\hat{\alpha}_{MLE})$ performed better than $\hat{p}_G(\hat{a}_{BRI})$ and 1P-BELP$(\hat{\alpha}_{BRI})$, only the results associated with the maximum likelihood approach are presented. Comparing Figure 2.1 to Figure 2.2 shows that these two aggregators make very good use of the training data and outperform the simple, parameterless aggregators as the training set increases from $K = 30$ to $K = 100$ problems. Overall, our aggregator $\hat{p}_G(\hat{a}_{MLE})$ achieves the lowest Brier score almost uniformly across Figures 2.2a to 2.2c. This result, however, is more of a sanity-check than a surprising result as the data were generated explicitly to match the model assumptions made by $\hat{p}_G$.

[4]This value was considered a realistic choice after analyzing the variance of the log-odds in our real-world data. The simulation was also run with unit variance. These results, however, were not remarkably different and are hence, for the sake of brevity, not presented in this paper




(a) (b) (c)

Figure 2.1: $K = 30$ synthetic problems for training. 1,000 problems for testing.




(a) (b) (c)

Figure 2.2: $K = 100$ synthetic problems for training. 1,000 problems for testing.

Based on Figures 2.1b and 2.2b, correcting for the bias when the data are actually unbiased ($a = 1.0$) does not cause much harm. But correcting for the bias when the data are truly biased ($a \neq 1.0$), yields noticeable performance benefits especially when $K = 100$. Interestingly, the mean performs better than all other aggregators when the experts are highly over-confident ($a \leq 0.7$) but is hugely outperformed when the experts

are under-confident ($a > 1$). To gain some understanding of this behavior, notice that in the highly over-confident case the distribution of the forecasts tends to be very skewed in the probability scale. The values in the long tail of such a distribution have a larger influence on the mean than, say, the median of the probability forecasts. Given that the median is mostly unaffected by these values, it produces an aggregate forecast that remains over-confident. The mean, by contrast, is drawn towards 0.5 by the values in the long tail. This produces an aggregate forecast that is less over-confident; hence improving forecast performance.

The improved performance, however, comes at a cost: when the true probability $p$ is very close to the extreme probabilities 0.0 and 1.0, the mean is, on average, the worst performer among all the aggregators in the analysis. This difference in performance, which is clear in Figures 2.1c and 2.2c, is more meaningful when compared to the baseline given by $p(1-p)^2 + (1-p)p^2$. Given that the expected Brier score is minimized at the true probability, this line should be considered as the ultimate goal in Figures 2.1c and 2.2c. Notice that all aggregators, except the mean, approach the line $p(1-p)^2 + (1-p)p^2$ from above as $p$ gets closer to the extreme probabilities 0.0 and 1.0.

### 2.3.2 Synthetic Data: Misspecified Model

Next we evaluate the different aggregators on data that have not been generated from the Logit-Normal distribution. The setup considered is an extension of the simulation study introduced in Ranjan and Gneiting (2010) and further applied in Allard et al. (2012). Under our extended version, the true probability for a problem with $N$ forecasters is given by

$$p \;=\; \Phi\left(\sum_{i=1}^{N} u_i\right),$$

where $\Phi$ is the cumulative distribution function of a standard normal distribution, and each $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Now, assume that the $i$th expert is aware of the true probability generating

process but only observes $u_i$. Then his calibrated estimate for $p$ is given by

$$p_i \;=\; \Phi\left(\frac{u_i}{\sqrt{2N-1}}\right)$$

Notice that the more forecasters are participating in a given problem, the less information (proportionally) knowing $u_i$ gives the forecaster. Therefore as the number of forecasters increases, the forecaster shrinks his estimate more and more towards 0.5. More specifically, $p_i \to \Phi(0) = 0.5$ for all $i = 1, \ldots, N$ as $N \to \infty$.

To give a real-world analogy of this setup, think of a group of $N$ people independently voting on a binary event. Knowing everybody's vote determines the final outcome. Given that each person only knows his own vote, his proportional knowledge share diminishes as more people enter the voting. As a result, his probability forecast for the final outcome should shrink towards 0.5.

In our simulation, we varied the number of forecasters per problem, $N$, from 2 to 100 (with increments of one forecaster). Under each value of $N$, the simulation ran for a total of 10,000 iterations. Each iteration produced the true probabilities for the $K$ problems and their associated pools of $N$ probability estimates from the process described above. The true probabilities were used to generate Bernoulli random variables that indicated which events occurred and which did not. Testing was performed on a separate testing set consisting of 1,000 problems, each with the same number of forecasters as the problems in the training set. In the end, the resulting Brier scores were averaged to give an average Brier score at each number of forecasters for each problem.

Figure 2.3 plots the average Brier score against the number of forecasters per problem under $K = 100$ problems. The same analysis was performed under $K = 30$. The results, however, turned out to be almost identical to the results under $K = 100$ and hence, for the sake of brevity, are not presented separately. Before discussing the $K = 100$ results, however, it is important to emphasize the peculiarity of this setting. Notice that, unlike in

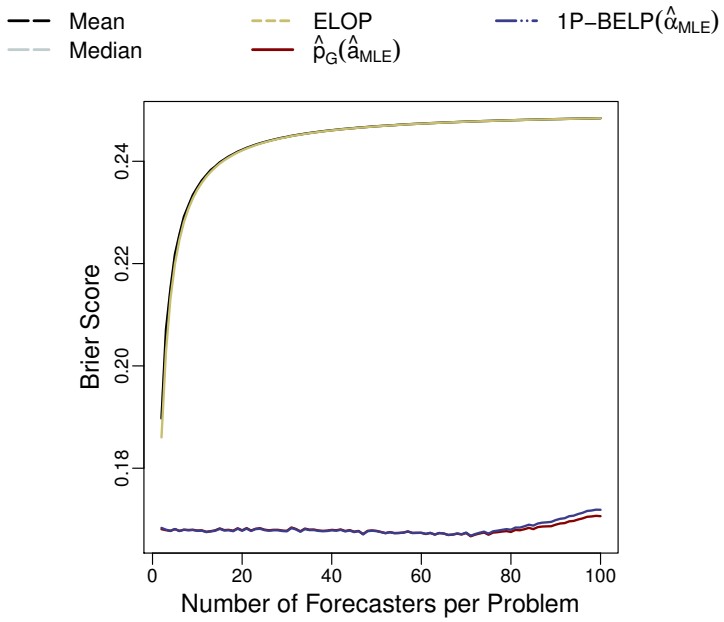



Figure 2.3: 100 synthetic problems for training. 1,000 problems for testing.

many generally encountered data generating processes, having more data leads to increased bias and is therefore harmful. As a result, we would expect the aggregators to perform worse as the sample size increases. Based on Figure 2.3, the mean, median, and ELOP, which do not aim to correct for the bias, in fact, degrade in terms of performance as the number of forecasters increases. The one-parameter aggregators, $\hat{p}_G$ and 1P-BELP, by contrast, are able to stabilize the average Brier score despite the increasing bias in the probability estimates. Overall, $\hat{p}_G$ achieves the lowest Brier score across all numbers of forecasters per problem.

### 2.3.3 Real Data: Predicting Geopolitical Events

We recruited over 1,300 forecasters, who ranged from graduate students to forecasting and political science faculty and practitioners, and then asked them to give probability forecasts on 69 geopolitical events. Forecasters were recruited from professional societies, research

centers, alumni associations, science bloggers, and word of mouth. Requirements included at least a Bachelor's degree and completion of psychological and political tests that took roughly two hours. These measures assessed cognitive styles, cognitive abilities, personality traits, political attitudes, and real-world knowledge. All forecasters knew that their probability estimates would be assessed for accuracy using Brier scores. This incentivized the forecasters to report their true beliefs instead of gaming the system. Forecasters received $150 for meeting minimum participation requirements, regardless of their accuracy. They also received status rewards for their performance via leaderboards displaying Brier scores for the top 20 forecasters. Each of the 69 geopolitical events had two possible outcomes. For instance, some of the questions were

> *Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?*

and

> *Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?*

See the Appendix for a complete list of the 69 problems and associated summary statistics.

The forecasters were allowed to update their forecast as long as the question was active. Some questions were active longer than others. The number of active days ranged from 2 to 173 days, with a mean of 54.7 days. It is important to note that this paper does not focus on dynamic data. Instead we study pools of probability forecasts with no more than one forecast given by a single expert. More specifically, we consider the first three days for each problem because this is when the most of uncertainty is present. If an expert made more than one forecast within these three days, we consider only his most recent forecast. This results in 69 sets of probabilities with the number of forecasters per problem ranging from 86 to 647 with a mean of 235.1 forecasters. Given that all experts did not participate

in every problem, we consider the experts completely anonymous (and interchangeable) both within and across problems. Before we evaluate the results, however, we discuss several practical matters that need to be taken into account when aggregating real-world forecasting data.

#### 2.3.3.1 Extreme Values and Inconsistent Data

For any value of $a$, the aggregator $\hat{p}_G$ satisfies the $0/1$ forcing property which states that if the pool of forecasts includes an extreme value, that is either zero or one but not both, then the estimator should return that extreme value (see, e.g., Allard et al. (2012)). This property is desirable if one of the forecasters happens to know the final outcome of the event with absolute confidence. Unfortunately, experts can make such absolute claims even when they are not completely sure of the outcome. For instance, each of the forecast pools associated with the 69 questions in our data contained both a zero and a one. In any such dataset, an aggregator that is based on the geometric mean of the odds is undefined.

These data inconsistencies can be avoided by adding and subtracting a small quantity from zeros and ones, respectively. Ariely et al. (2000) suggest changing $p = 0$ and $1$ to $p = 0.02$ and $0.98$, respectively. Allard et al. (2012) only consider probabilities that fall within a constrained interval, say $[0.001, 0.999]$, and throw out the rest. Given that this implies ignoring a portion of the data, we take an approach similar to that of Ariely et al. (2000) and replace $p = 0$ and $1$ with $p = 0.01$ and $0.99$, respectively. In the case of multinomial events, the modified probabilities should be normalized to sum to one. This forces the probability estimates to the open interval $(0, 1)$. The transformation will shift the truncated values even closer to their true extreme values. For instance, if $a$ is larger than two, which often is the case, 0.01 and 0.99 would be transformed at least to 0.0001 and 0.9999, respectively.

Another practical solution is to estimate the geometric mean of the odds with the odds

given by the arithmetic mean of the probabilities. This gives us the following estimator

$$\hat{p}_A(a) \quad = \quad \frac{\left[\frac{\bar{p}}{1-\bar{p}}\right]^a}{1+\left[\frac{\bar{p}}{1-\bar{p}}\right]^a},$$

where $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$. The subindex emphasizes the use of the arithmetic mean. The two estimators $\hat{p}_G$ and $\hat{p}_A$ will differ the most when the set of probability forecasts includes values close to the extremes. Therefore the larger the variance term $\sigma^2$ of the Logit-Normal model is the more we would expect these two estimators to differ. For comparison's sake, we have included $\hat{p}_A$ in the real-world data analysis.

A similar problem arises with the logarithmic opinion pool, where zero predictions from experts can be viewed as "vetoes" (see Genest and Zidek (1986)). To address this, we replaced $p = 0$ with $p = 0.01$ and normalized the probabilities to sum to one.

#### 2.3.3.2 Aggregator Comparison on Expert Data

This section evaluates the aggregators on the first three days of the 69 problems in our dataset. The evaluation begins by exploring the impact of number of forecasters per problem on predictive power. Each run of the simulation fixes the number of forecasters per problem and samples a random subset (of this size) of forecasters within each problem. These subsets are then used for training and computing a Brier score. The sampling procedure is repeated 1,000 times during the simulation. The resulting 1,000 Brier scores are averaged to obtain an overall performance measure under the given number of forecasters per problem.

Figure 2.4 plots the average Brier score against the number of forecasters per problem. The MLE aggregator $\hat{p}_G$ achieves the lowest Brier score across all numbers of forecasters per problem. The two aggregators $\hat{p}_A$ and P1-BELP perform so similarly that their average Brier scores are almost indistinguishable. The performance gap from $\hat{p}_G$ to P1-BELP and

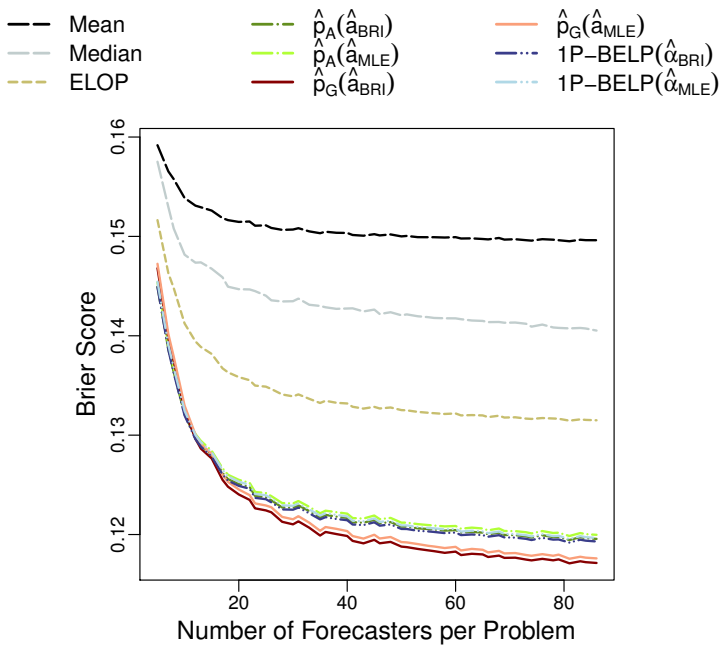



Figure 2.4: 69 real-world problems for training. The first three days; much uncertainty.

$\hat{p}_A$ appears to widen as the number of forecasters increases.

It merits note that most of the improvement across the different approaches occurs before roughly 50 forecasters per problem. This suggests a new strategy for collecting data: instead of having a large number of forecasters making predictions on a few problems, we should have around 50 forecasters involved with a large number of problems. With a larger number of problems, a more accurate estimate of the systematic bias could be acquired, possibly leading to improved forecast performance.

A similar analysis was performed with the last three days of each problem. The average Brier scores, however, were very close to zero. There was, in fact, so much certainty among the forecasters, that simply taking the median gave an aggregate forecast very close to the truth. For this reason, we decided to not present these results in this paper.

The average Brier scores in Figure 2.4 are based on the training error. No separate testing set was used in this particular analysis because we believe that fitting one parameter

in a large enough sample does not overfit significantly. Figure 2.5 plots the Brier score for $\hat{p}_G$ under varying levels of $a$. Given that the optimal level of $a$ is around 2.0, the experts (as a group) appear under confident, and $\hat{p}_G$ gains its advantage by shifting each of the probability forecasts closer to its nearest boundary point (0.0 or 1.0).

Running a *repeated sub-sampling validation* with a training set of size $K$ and a testing set of size $69-K$ supports the results shown in Figure 2.4. Table 2.1 shows the results after running *repeated sub-sampling validation* with $K = 30$ and $K = 60$ a total of 1,000 times and then averaging the resulting 1,000 (testing) Brier scores. For the sake of consistency, we have also included the average logarithmic scores:

$$-\frac{1}{69-K}\sum_{k=1}^{69-K} Z_k \log(\hat{p}_k) + (1-Z_k)\log(1-\hat{p}_k)$$

where $\hat{p}_j$ is the probability estimate and $Z_j$ is the event indicator for the $j$th testing problem defined earlier in Section 2.2.2. The values given in parentheses are the estimated standard deviations of the testing scores. Given that the mean, median, and ELOP do not use training data, their reported scores are based on the simulation with $K = 30$ that uses a larger testing set.

In Table 2.1 we have also included the bias-corrected versions of the mean, median, and ELOP. This correction was attained by applying bootstrap sampling to the pool of probabilities for a total of 1,000 times. More specifically,

$$\hat{p}_{f,k} \;=\; 2f(\boldsymbol{p}_k) - \frac{1}{1000}\sum_{i=1}^{1000} f\left(\boldsymbol{p}_{k,bs}^{(i)}\right),$$

where $\boldsymbol{p}_k$ is the (full) original set of probabilities for the $k$th problem, $\boldsymbol{p}_{k,bs}^{(i)}$ is the $i$th bootstrap sample obtained from $\boldsymbol{p}_k$, and $f$ is a functional depending on the estimator. For instance, when correcting the sample median, $f(\boldsymbol{p}_k) = \text{median}(\boldsymbol{p}_k)$. The biases found, however, turned out to be very small. As can be seen in Table 2.1, correcting for the bias

| | Brier Score | | Logarithmic Score | |
|---|---|---|---|---|
| | Bias Correction | | Bias Correction | |
| | No | Yes | No | Yes |
| Mean | 0.150 (0.032) | 0.150 (0.032) | 0.477 (0.025) | 0.477 (0.025) |
| Median | 0.140 (0.038) | 0.139 (0.038) | 0.446 (0.031) | 0.444 (0.031) |
| ELOP | **0.132 (0.039)** | **0.131 (0.039)** | **0.425 (0.032)** | **0.425 (0.032)** |
| | $K = 30$ | | | |
| | Bias Estimation | | Bias Estimation | |
| | BRI | MLE | BRI | MLE |
| 1P-BELP | 0.126 (0.027) | 0.125 (0.026) | 0.401 (0.115) | 0.401 (0.117) |
| $\hat{p}_A$ | 0.127 (0.027) | 0.125 (0.026) | 0.402 (0.109) | 0.402 (0.115) |
| $\hat{p}_G$ | **0.124 (0.028)** | **0.122 (0.026)** | **0.401 (0.134)** | **0.394 (0.127)** |
| | $K = 60$ | | | |
| | Bias Estimation | | Bias Estimation | |
| | BRI | MLE | BRI | MLE |
| 1P-BELP | 0.122 (0.061) | 0.121 (0.064) | 0.383 (0.170) | 0.384 (0.193) |
| $\hat{p}_A$ | 0.122 (0.061) | 0.121 (0.065) | 0.385 (0.168) | 0.386 (0.190) |
| $\hat{p}_G$ | **0.119 (0.060)** | **0.118 (0.064)** | **0.376 (0.165)** | **0.377 (0.188)** |

Table 2.1: $K$ problems for training. $69 - K$ problems for testing. 1,000 repetitions. The values in the parentheses are the estimated standard deviations of the testing scores.

improved the performance only by a small margin if at all.

For convenience, we have bolded the lowest scores in each column of the three boxes. Overall, the ranking of the aggregators on relative performances is the same as in Figure 2.4. As seen before, $\hat{p}_G(\hat{a}_{MLE})$ achieves the lowest Brier and logarithmic scores by a noticeable margin.

#### 2.3.3.3 Less Transformation for More Expertise

Earlier we proposed that the more expertise the forecaster has, the less systematic bias can be found in his probability forecasts. This means that his forecasts require less transformation, i.e. a lower level of $a$. To evaluate this interpretation, we asked forecasters to self-assess their level of expertise on the topic. The level of expertise was measured on a 1-to-5 scale (1 = Not At All Expert to 5 = Extremely Expert). Figure 2.6 plots the max-

imum likelihood estimator of $a$ under different levels of expertise. The gray bars around each point are the 95% (Bonferroni corrected) simultaneous confidence intervals computed by inverting the likelihood-ratio test. We have allowed for values of $a$ less than 1 to reveal possible overconfidence.

These results are based on the first three days of data for each problem because this is when most uncertainty is present and the expertise level matters the most. Although we are unable to show statistical significance for a strictly decreasing trend in the systematic bias across the different levels of expertise, the need for transformation is apparent in a 99% confidence interval for the value of $a$ when the level of expertise is not taken into account. This interval (not shown on Figure 2.6) is $[1.161, 3.921]$. Given that it does not include $a = 1$, i.e. the level of no transformation, there is significant evidence (at 1% significance level) that as a group the experts are under-confident. Therefore their probability forecasts should be shifted more toward the extreme probabilities 0.0 and 1.0.

## 2.4 Conclusions

In this paper we have motivated and derived a model-based approach to aggregation of expert probability forecasts. The resulting aggregator, which is based on the geometric mean of the expert odds, has a single tuning parameter that determines how much each of the probabilities should be shifted toward its nearest extreme probability 0.0 or 1.0. This transformation aims to compensate for the under-confidence that arises from incomplete knowledge and becomes present at a group level among the experts. That is, although the individual experts may not all be under-confident (in fact, according to our analysis, some of the experts with high self-reported expertise tend to be over-confident), as a group the experts are under-confident. Therefore, if no bias-correction is performed, the consensus probability forecast can turn out to be biased and sub-optimal in terms of forecast performance.

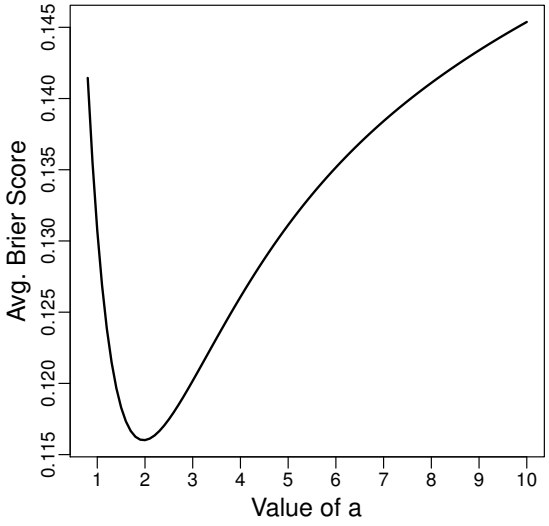



Figure 2.5: Sensitivity to the choice of $a$ based on all data available in the first three days.

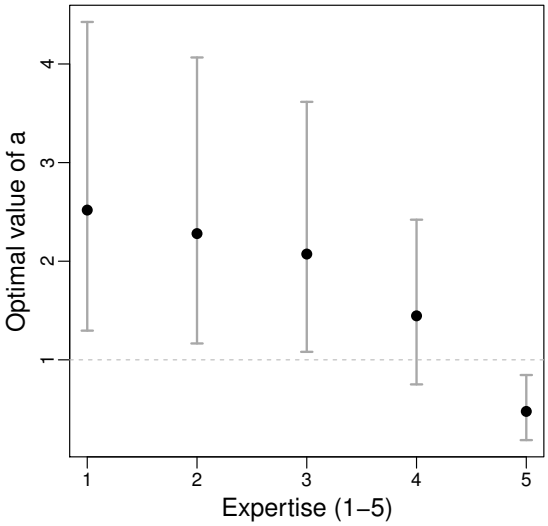



Figure 2.6: Optimal transformation $a$ (representing systematic bias) with 95% simultaneous confidence intervals as a function of forecaster self-reported expertise.

To study the extent of this bias, it is helpful to compare the aggregate probability forecasts given by a naive approach, such as the arithmetic mean with no explicit bias-correction, against the corresponding forecasts given by a bias-correcting approach, such as our $\hat{p}_G(\hat{a}_{MLE})$ aggregator. In the table in the Appendix we have provided both the mean probability forecast and the aggregate estimate $\hat{p}_G(\hat{a}_{MLE})$ for all the 69 problems in our real-world dataset. Looking at these estimates (see Figure A.1 in the Appendix), it is clear that the $\hat{p}_G(\hat{a}_{MLE})$ aggregator is much sharper than the simple arithmetic mean. Furthermore, the noticeable disagreement between the two estimates (with mean absolute difference of 0.175) suggests that a large enough bias persists for bias-correction to improve performance.

As is evident throughout Section 2.3, our aggregator shows very good forecast performance especially when the outcome of the event involves much uncertainty. In addition,

our aggregator utilizes the training data efficiently leading to improved forecast performance as the size of the training set increases. This improvement, however, happens at such a diminishing rate that there are few additional gains in forecast performance from aggregating more than roughly 50 forecasters per problem (see Figure 2.4).

It is likely that our aggregator can be improved and extended in many ways. This, however, might lead to reduced interpretability and additional assumptions that may not comply with the psychology literature. For instance, being able to estimate the bias term $a$ within each problem individually could improve the performance of the aggregator. This, however, seems difficult given the framework of this paper; that is, non-dynamic probability pools given by interchangeable forecasters. As Table 2.1 shows, simple bootstrap approaches to problem-specific bias-correction do not lead to significant improvements in forecast performance.

Perhaps an intermediate approach that neither shares a single bias term nor has completely independent bias terms across problems will yields further improvements in performance. One possibility is that the more difficult the problem, the more bias persists among the experts. This suggests that developing a measure of the difficulty of the problem, estimating a single bias term across all problems, and then adjusting this bias term individually for each problem based on the estimated difficulty could lead to better predictions. Coming up with a reasonable difficulty measure, however, is challenging. One simple idea is to use the variance of the expert forecasts as a proxy for problem difficulty.

Such an extension could also satisfy the unanimity property: if all experts give the same forecast, then the aggregator should return that forecast as a unanimous decision. Although this property may not be critical in large probability pools such our dataset, it needs to be mentioned that our aggregator does not satisfy the unanimity property. Instead, it tends to assume under-confidence and shift each of the probability forecasts closer to its nearest extreme probability 0.0 or 1.0. Nonetheless, it gives extremely good results on

real data. Furthermore, unlike the Beta-transformed linear opinion pool, our aggregator can be applied to a wide range of situations, such as events with more than two possible outcomes, and has a simple analytic form making it interpretable, flexible, and amenable to many future extensions.

## 2.5 Acknowledgements

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. We deeply appreciate the project management skills and work of Terry Murray and David Wayrynen, which went far beyond the call-of-duty on this project.

# 3

# Probability Aggregation in Time-Series: Dynamic Hierarchical Modeling of Sparse Expert Beliefs*


## Abstract

Most subjective probability aggregation procedures use a single probability judgment from each expert, even though it is common for experts studying real problems to update their probability estimates over time. This paper advances into unexplored areas of probability aggregation by considering a dynamic context in which experts can update their beliefs at random intervals. The updates occur very infrequently, resulting in a sparse dataset that cannot be modeled by standard time-series procedures. In response to the lack of appropriate methodology, this paper presents a hierarchical model that takes into account the expert's level of self-reported expertise and produces aggregate probabilities that are sharp and well-calibrated both in- and out-of-sample. The model is demonstrated on a real-world dataset that includes over 2,300 experts making multiple probability forecasts over two years on different subsets of 166 international political events.


*Joint work with Shane T. Jensen, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar

## 3.1 Introduction

Experts' probability assessments are often evaluated on *calibration*, which measures how closely the frequency of event occurrence agrees with the assigned probabilities. For instance, consider all events that an expert believes to occur with a 60% probability. If the expert is well-calibrated, 60% of these events will actually end up occurring. Even though several experiments have shown that experts are often poorly calibrated (see, e.g., Cooke (1991); Shlyakhter et al. (1994)), these are noteworthy exceptions. In particular, Wright et al. (1994) argue that higher self-reported expertise can be associated with better calibration.

Calibration by itself, however, is not sufficient for useful probability estimation. Consider a relatively stationary process, such as rain on different days in a given geographic region, where the observed frequency of occurrence in the last 10 years is 45%. In this setting an expert could always assign a constant probability of 0.45 and be well-calibrated. His assessment, however, can be made without any subject-matter expertise. For this reason the long-term frequency is often considered the baseline probability – a naive assessment that provides the decision-maker very little extra information. Experts should make probability assessments that are as far from the baseline as possible. The extent to which their probabilities differ from the baseline is measured by *sharpness* (Gneiting et al. (2008); Winkler and Jose (2008)). If the experts are both sharp and well-calibrated, they can forecast the behavior of the process with high certainty and accuracy. Therefore useful probability estimation should maximize sharpness subject to calibration (see, e.g., Raftery et al. (2005); Murphy and Winkler (1987b)).

There is strong empirical evidence that bringing together the strengths of different experts by combining their probability forecasts into a single consensus, known as the *crowd belief*, improves predictive performance. Prompted by the many applications of probability forecasts, including medical diagnosis (Wilson et al. (1998); Pepe (2003)), political and

socio-economic foresight (Tetlock (2005)), and meteorology (Sanders (1963); Vislocky and Fritsch (1995); Baars and Mass (2005)), researchers have proposed many approaches to combining probability forecasts (see, e.g., Ranjan and Gneiting (2010); Satopää et al. (2014); Batchelder et al. (2010) for some recent studies, and Genest and Zidek (1986); Wallsten et al. (1997); Clemen and Winkler (2007); Primo et al. (2009) for a comprehensive overview). The general focus, however, has been on developing one-time aggregation procedures that consult the experts' advice only once before the event resolves.

Consequently, many areas of probability aggregation still remain rather unexplored. For instance, consider investors aiming to assess whether a stock index will finish trading above a threshold on a given date. To maximize their overall predictive accuracy, they may consult a group of experts repeatedly over a period of time and adjust their estimate of the aggregate probability accordingly. Given that the experts are allowed to update their probability assessments, the aggregation should be performed by taking into account the temporal correlation in their advice.

This paper adds another layer of complexity by assuming a heterogeneous set of experts, most of whom only make one or two probability assessments over the hundred or so days before the event resolves. This means that the decision-maker faces a different group of experts every day, with only a few experts returning later on for a second round of advice. The problem at hand is therefore strikingly different from many time-series estimation problems, where one has an observation at every time point – or almost every time point. As a result, standard time-series procedures like ARIMA (see, e.g., Mills (1991)) are not directly applicable. This paper introduces a time-series model that incorporates self-reported expertise and captures a sharp and well-calibrated estimate of the crowd belief. The model is highly interpretable and can be used for:

- analyzing under- and overconfidence in different groups of experts,
- obtaining accurate probability forecasts, and

Table 3.1: Five-number summaries of our real-world data.

| Statistic | Min. | $Q_1$ | Median | Mean | $Q_3$ | Max. |
|---|---|---|---|---|---|---|
| # of Days a Question is Active | 4 | 35.6 | 72.0 | 106.3 | 145.20 | 418 |
| # of Experts per Question | 212 | 543.2 | 693.5 | 783.7 | 983.2 | 1690 |
| # Forecasts given by each Expert on a Question | 1 | 1.0 | 1.0 | 1.8 | 2.0 | 131 |
| # Questions participated by an Expert | 1 | 14.0 | 36.0 | 55.0 | 90.0 | 166 |

Table 3.2: Frequencies of the self-reported expertise (1 = Not At All Expert and 5 = Extremely Expert) levels across all the 166 questions in our real-world data.

| Expertise Level | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Frequency (%) | 25.3 | 30.7 | 33.6 | 8.2 | 2.1 |

- gaining question-specific quantities with easy interpretations, such as expert disagreement and problem difficulty.

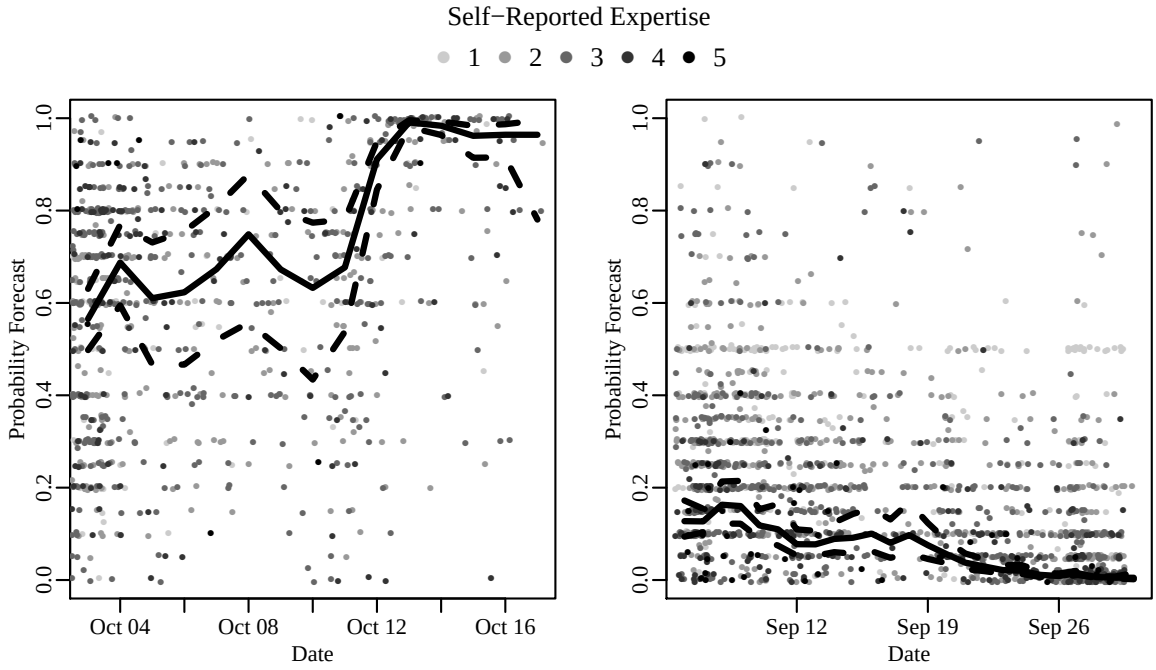
This paper begins by describing our geopolitical database. It then introduces a dynamic hierarchical model for capturing the crowd belief. The model is estimated in a two-step procedure: first, a sampling step produces constrained parameter estimates via Gibbs sampling (see, e.g., Geman and Geman (1984)); second, a calibration step transforms these estimates to their unconstrained equivalents via a one-dimensional optimization procedure. The model introduction is followed by the first evaluation section that uses synthetic data to study how accurately the two-step procedure can estimate the crowd belief. The second evaluation section applies the model to our real-world geopolitical forecasting database. The paper concludes with a discussion of future research directions and model limitations.

## 3.2 Geopolitical Forecasting Data

Forecasters were recruited from professional societies, research centers, alumni associations, science bloggers, and word of mouth ($n = 2,365$). Requirements included at least a Bachelor's degree and completion of psychological and political tests that took roughly two hours. These measures assessed cognitive styles, cognitive abilities, personality traits, political attitudes, and real-world knowledge. The experts were asked to give probability forecasts (to the second decimal point) and to self-assess their level of expertise (on a 1-to-5 scale with 1 = Not At All Expert and 5 = Extremely Expert) on a number of 166 geopolitical binary events taking place between September 29, 2011 and May 8, 2013. Each question was active for a period during which the participating experts could update their forecasts as frequently as they liked without penalty. The experts knew that their probability estimates would be assessed for accuracy using Brier scores[2]. This incentivized them to report their true beliefs instead of attempting to game the system (Winkler and Murphy (1968)). In addition to receiving $150 for meeting minimum participation requirements that did not depend on prediction accuracy, the experts received status rewards for their performance via leader-boards displaying Brier scores for the top 20 experts. Given that a typical expert participated only in a small subset of the 166 questions, the experts are considered indistinguishable conditional on the level of self-reported expertise.

The average number of forecasts made by a single expert in one day was around 0.017, and the average group-level response rate was around 13.5 forecasts per day. Given that the group of experts is large and diverse, the resulting dataset is very sparse. Tables 3.1 and 3.2 provide relevant summary statistics on the data. Notice that the distribution of the self-reported expertise is skewed to the right and that some questions remained active longer than others. For more details on the dataset and its collection see Ungar et al. (2012) and

[2]The Brier score is the squared distance between the probability forecast and the event indicator that equals 1.0 or 0.0 depending on whether the event happened or not, respectively. See Brier (1950) for the original introduction.




(a) Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?

(b) Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?

Figure 3.1: Scatterplots of the probability forecasts given for two questions in our dataset. The solid line gives the posterior mean of the calibrated crowd belief as estimated by our model. The surrounding dashed lines connect the point-wise 95% posterior intervals.

Satopää et al. (2014).

To illustrate the data with some concrete examples, Figures 3.1a and 3.1b show scatterplots of the probability forecasts given for (a) *Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011?*, and (b) *Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011?*. The points have been shaded according to the level of self-reported expertise and jittered slightly to make overlaps visible. The solid line gives the posterior mean of the calibrated crowd belief as estimated by our model. The surrounding dashed lines connect the point-wise 95% posterior intervals. Given that the European bailout fund was ratified before November 1, 2011 and

that the Nikkei 225 index finished trading at around 8,700 on September 30, 2011, the general trend of the probability forecasts tends to converge towards the correct answers. The individual experts, however, sometimes disagree strongly, with the disagreement persisting even near the closing dates of the questions.

## 3.3 Model

Let $p_{i,t,k} \in (0,1)$ be the probability forecast given by the $i$th expert at time $t$ for the $k$th question, where $i = 1,\ldots,I_k$, $t = 1,\ldots,T_k$, and $k = 1,\ldots,K$. Denote the logit-probabilities with

$$Y_{i,t,k} = \text{logit}(p_{i,t,k}) = \log\left(\frac{p_{i,t,k}}{1-p_{i,t,k}}\right) \in \mathbb{R}$$

and collect the logit-probabilities for question $k$ at time $t$ into a vector

$$\boldsymbol{Y}_{t,k} = [Y_{1,t,k}\ Y_{2,t,k}\ \ldots\ Y_{I_k,t,k}]^T.$$

Partition the experts into $J$ groups based on some individual feature, such as self-reported expertise, with each group sharing a common multiplicative bias term $b_j \in \mathbb{R}$ for $j = 1,\ldots,J$. Collect these bias terms into a bias vector $\boldsymbol{b} = [b_1\ b_2\ \ldots\ b_J]^T$. Let $\boldsymbol{M}_k$ be a $I_k \times J$ matrix denoting the group-memberships of the experts in question $k$; that is, if the $i$th expert participating in the $k$th question belongs to the $j$th group, then the $i$th row of $\boldsymbol{M}_k$ is the $j$th standard basis vector $\boldsymbol{e}_j$. The bias vector $\boldsymbol{b}$ is assumed to be identical across all $K$ questions. Under this notation, the model for the $k$th question can be expressed as

$$\boldsymbol{Y}_{t,k} = \boldsymbol{M}_k\boldsymbol{b}X_{t,k} + \boldsymbol{v}_{t,k} \tag{3.1}$$

$$X_{t,k} = \gamma_k X_{t-1,k} + w_{t,k} \tag{3.2}$$

$$X_{0,k} \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

where (3.1) denotes the observed process, (3.2) shows the hidden process that is driven by the constant $\gamma_k \in \mathbb{R}$, and $(\mu_0, \sigma_0^2) \in (\mathbb{R}, \mathbb{R}^+)$ are hyper-parameters fixed *a priori* to 0 and 1, respectively. The error terms follow

$$\begin{aligned} \boldsymbol{v}_{t,k}|\sigma_k^2 &\overset{i.i.d.}{\sim} \mathcal{N}_{I_k}(\mathbf{0}, \sigma_k^2 \boldsymbol{I}_{I_k}) \\ w_{t,k}|\tau_k^2 &\overset{i.i.d.}{\sim} \mathcal{N}(0, \tau_k^2), \end{aligned}$$

Therefore the parameters of the model are $\boldsymbol{b}$, $\sigma_k^2$, $\gamma_k$, and $\tau_k^2$ for $k = 1, \ldots, K$. Their prior distributions are chosen to be non-informative, $p(\boldsymbol{b}, \sigma_k^2|\boldsymbol{X}_k) \propto \sigma_k^2$ and $p(\gamma_k, \tau_k^2|\boldsymbol{X}_k) \propto \tau_k^2$.

The hidden state $X_{t,k}$ represents the aggregate logit-probability for the $k$th event given all the information available up to and including time $t$. To make this more specific, let $Z_k \in \{0, 1\}$ indicate whether the event associated with the $k$th question happened $(Z_k = 1)$ or did not happen $(Z_k = 0)$. If $\{\mathcal{F}_{t,k}\}_{t=1}^{T_k}$ is a filtration representing the information available up to and including a given time point, then according to our model $\mathbb{E}[Z_k|\mathcal{F}_{t,k}] = \mathbb{P}(Z_k = 1|\mathcal{F}_{t,k}) = \text{logit}^{-1}(X_{t,k})$. Ideally this probability maximizes sharpness subject to calibration (for technical definitions of calibration and sharpness see Ranjan and Gneiting (2010); Gneiting and Ranjan (2013)) Even though a single expert is unlikely to have access to all the available information, a large and diverse group of experts may share a considerable portion of the available information. The collective wisdom of the group therefore provides an attractive proxy for $\mathcal{F}_{t,k}$.

Given that the experts may believe in false information, hide their true beliefs, or be biased for many other reasons, their probability assessments should be aggregated via a model that can detect potential bias, separate signal from noise, and use the collective opinion to estimate $X_{t,k}$. In our model the experts are assumed to be, on average, a multiplicative constant $\boldsymbol{b}$ away from $X_{t,k}$. Therefore an individual element of $\boldsymbol{b}$ can be interpreted as a

group-specific *systematic bias* that labels the group either as over-confident ($b_j \in (1, \infty)$) or as under-confident ($b_j \in (0, 1)$). See Section 3.3 for a brief discussion on different bias structures. Any other deviation from $X_{t,k}$ is considered *random noise*. This noise is measured in terms of $\sigma_k^2$ and can be assumed to be caused by momentary over-optimism (or pessimism), false beliefs, or other misconceptions.

The *random fluctuations* in the hidden process are measured by $\tau_k^2$ and are assumed to represent changes or shocks to the underlying circumstances that ultimately decide the outcome of the event. The *systematic component* $\gamma_k$ allows the model to incorporate a constant signal stream that drifts the hidden process. If the uncertainty in the question diminishes ($\gamma_k \in (1, \infty)$), the hidden process drifts to positive or negative infinity. Alternatively, the hidden process can drift to zero in which case any available information does not improve predictive accuracy ($\gamma_k \in (0, 1)$). Given that all the questions in our dataset were resolved within a pre-specified timeframe, we expect $\gamma_k \in (1, \infty)$ for all $k = 1, \ldots, K$.

As for any future time $T^* \geq t$

$$\begin{aligned} X_{T^*,k} &= \gamma_k^{T^*-t} X_t + \sum_{i=t+1}^{T^*} \gamma_k^{T^*-i} w_i \\ &\sim \mathcal{N}\left(\gamma_k^{T^*-t} X_{t,k}, \tau_k^2 \sum_{i=t+1}^{T^*} \gamma_k^{T^*-i}\right), \end{aligned}$$

the model can be used for time-forward prediction as well. The prediction for the aggregate logit-probability at time $T^*$ is given by an estimate of $\gamma^{T^*-t} X_{t,k}$. Naturally the uncertainty in this prediction grows in $T$. To make such time-forward predictions it is necessary to assume that the past population of experts is representative of the future population. This is a reasonable assumption because even though the future population may consist of entirely different individuals, on average the population is likely to look very similar to the past population. In practice, however, social scientists are generally more interested in an estimate of the current probability than the probability under unknown conditions in the future.

For this reason, our analysis focuses on probability aggregation only up to the current time $t$.

For the sake of model identifiability, it is sufficient to share only one of the elements of $\boldsymbol{b}$ among the $K$ questions. In this paper, however, all the elements of $\boldsymbol{b}$ are assumed to be identical across the questions because some of the questions in our real-world data set involve very few experts with the highest level of self-reported expertise. The model can be extended rather easily to estimate bias at a more general level. For instance, by assuming a hierarchical structure $b_{ik} \sim \mathcal{N}\left(b_{j(i,k)}, \sigma^2_{j(i,k)}\right)$, where $j(i,k)$ denotes the self-reported expertise of the $i$th expert in question $k$, the bias can be estimated at an individual-level. These estimates can then be compared across questions. Individual-level analysis was not performed in our analysis for two reasons. First, most experts gave only a single prediction per problem, which makes accurate bias estimation at the individual-level very difficult. Second, it is unclear how the individually estimated bias terms can be validated.

If the future event can take upon $M > 2$ possible outcomes, the hidden state $X_{t,k}$ is extended to a vector of size $M-1$ and one of the outcomes, e.g., the $M$th one, is chosen as the base-case to ensure that the probabilities will sum to one at any given time point. Each of the remaining $M-1$ possible outcomes is represented by an observed process similar to (3.1). Given that this multinomial extension is equivalent to having $M-1$ independent binary-outcome models, the estimation and properties of the model are easily extended to the multi-outcome case. This paper focuses on binary-outcomes because it is the most commonly encountered setting in practice.

## 3.4 Model Estimation

This section introduces a two-step procedure, called *Sample-And-Calibrate* (SAC), that captures a well-calibrated estimate of the hidden process without sacrificing the interpretability of our model.

### 3.4.1 Sampling Step

Given that $(a\boldsymbol{b}, X_{t,k}/a, a^2\tau_k^2) \neq (\boldsymbol{b}, X_{t,k}, \tau_k^2)$ for any $a > 0$ yield the same likelihood for $\boldsymbol{Y}_{t,k}$, the model as described by (3.1) and (3.2) is not identifiable. A well-known solution is to choose one of the elements of $\boldsymbol{b}$, say $b_3$, as the reference point and fix $b_3 = 1$. In Section 3.5 we provide a guideline for choosing the reference point. Denote the constrained version of the model by

$$\begin{aligned}
\boldsymbol{Y}_{t,k} &= \boldsymbol{M}_k \boldsymbol{b}(1) X_{t,k}(1) + \boldsymbol{v}_{t,k} \\
X_{t,k}(1) &= \gamma_k(1) X_{t-1,k}(1) + w_{t,k} \\
\boldsymbol{v}_{t,k} | \sigma_k^2(1) &\overset{i.i.d.}{\sim} \mathcal{N}_{I_k}(\boldsymbol{0}, \sigma_k^2(1) \boldsymbol{I}_{I_k}) \\
w_{t,k} | \tau_k^2(1) &\overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tau_k^2(1)\right),
\end{aligned}$$

where the trailing input notation, $(a)$, signifies the value under the constraint $b_3 = a$. Given that this version is identifiable, estimates of the model parameters can be obtained. Denote the estimates by placing a hat on the parameter symbol. For instance, $\hat{\boldsymbol{b}}(1)$ and $\hat{X}_{t,k}(1)$ represent the estimates of $\boldsymbol{b}(1)$ and $X_{t,k}(1)$, respectively.

These estimates are obtained by first computing a posterior sample via Gibbs sampling and then taking the average of the posterior sample. The first step of our Gibbs sampler is to sample the hidden states via the *Forward-Filtering-Backward-Sampling* (FFBS) algorithm. FFBS first predicts the hidden states using a Kalman filter and then performs a backward sampling procedure that treats these predicted states as additional observations (see, e.g., Carter and Kohn (1994); Migon et al. (2005) for details on FFBS). Given that the Kalman filter can handle varying numbers or even no forecasts at different time points, it plays a very crucial role in our probability aggregation under sparse data.

Our implementation of the sampling step is written in C++ and runs quite quickly. To obtain 1000 posterior samples for 50 questions each with 100 time points and 50 experts

takes about 215 seconds on a 1.7 GHz Intel Core i5 computer. See the supplemental article for the technical details of the sampling steps (Satopää et al. (2014)), and, e.g., Gelman et al. (2003) for a discussion on the general principles of Gibbs sampling.

### 3.4.2 Calibration Step

Given that the model parameters can be estimated by fixing $b_3$ to any constant, the next step is to search for the constant that gives an optimally sharp and calibrated estimate of the hidden process. This section introduces an efficient procedure that finds the optimal constant without requiring any additional runs of the sampling step. First, assume that parameter estimates $\hat{\boldsymbol{b}}(1)$ and $\hat{X}_{t,k}(1)$ have already been obtained via the sampling step described in Section 3.4.1. Given that for any $\beta \in \mathbb{R}/\{0\}$,

$$\begin{aligned} \boldsymbol{Y}_{t,k} &= \boldsymbol{M}_k \boldsymbol{b}(1) X_{t,k}(1) + \boldsymbol{v}_{t,k} \\ &= \boldsymbol{M}_k \left(\boldsymbol{b}(1)\beta\right) \left(X_{t,k}(1)/\beta\right) + \boldsymbol{v}_{t,k} \\ &= \boldsymbol{M}_k \boldsymbol{b}(\beta) X_{t,k}(\beta) + \boldsymbol{v}_{t,k}, \end{aligned}$$

we have that $\boldsymbol{b}(\beta) = \boldsymbol{b}(1)\beta$ and $X_{t,k}(\beta) = X_{t,k}(1)/\beta$. Recall that the hidden process $X_{t,k}$ is assumed to be sharp and well-calibrated. Therefore $b_3$ can be estimated with the value of $\beta$ that simultaneously maximizes the sharpness and calibration of $\hat{X}_{t,k}(1)/\beta$. A natural criterion for this maximization is given by the class of *proper scoring rules* that combine sharpness and calibration (Gneiting et al. (2008); Buja et al. (2005)). Due to the possibility of *complete separation* in any one question (see, e.g., Gelman et al. (2008)), the maximization must be performed over multiple questions. Therefore,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}/\{0\}}{\arg\max} \sum_{k=1}^{K} \sum_{t=1}^{T_k} S\left(Z_k, \hat{X}_{k,t}(1)/\beta\right) \tag{3.3}$$

where $Z_k \in \{0, 1\}$ is the event indicator for question $k$. The function $S$ is a strictly proper scoring rule such as the negative Brier score (Brier (1950))

$$S_{BRI}(Z, X) \;=\; -(Z - \text{logit}^{-1}(X))^2$$

or the logarithmic score (Good (1952))

$$S_{LOG}(Z, X) \;=\; Z \log\left(\text{logit}^{-1}(X)\right) + (1 - Z) \log\left(1 - \text{logit}^{-1}(X)\right)$$

The estimates of the unconstrained model parameters are then given by

$$\begin{aligned}
\hat{X}_{t,k} &= \hat{X}_{k,t}(1)/\hat{\beta} \\
\hat{\boldsymbol{b}} &= \hat{\boldsymbol{b}}(1)\hat{\beta} \\
\hat{\tau}_k^2 &= \hat{\tau}_k^2(1)/\hat{\beta}^2 \\
\hat{\sigma}_k^2 &= \hat{\sigma}_k^2(1) \\
\hat{\gamma}_k &= \hat{\gamma}_k(1)
\end{aligned}$$

Notice that estimates of $\sigma_k^2$ and $\gamma_k$ are not affected by the constraint.

## 3.5 Synthetic Data Results

This section uses synthetic data to evaluate how accurately the SAC-procedure captures the hidden states and bias vector. The hidden process is generated from standard Brownian motion. More specifically, if $Z_{t,k}$ denotes the value of a path at time $t$, then

$$\begin{aligned}
Z_k &= \mathbf{1}\left(Z_{T_k,k} > 0\right) \\
X_{t,k} &= \text{logit}\left[\Phi\left(\frac{Z_{t,k}}{\sqrt{T_k - t}}\right)\right]
\end{aligned}$$

gives a sequence of $T_k$ calibrated logit-probabilities for the event $Z_k = 1$. A hidden process is generated for $K$ questions with a time horizon of $T_k = 101$. The questions involve 50 experts allocated evenly among five expertise groups. Each expert gives one probability forecast per day with the exception of time $t = 101$ when the event resolves. The forecasts are generated by applying bias and noise to the hidden process as described by (3.1). Our simulation study considers a three-dimensional grid of parameter values:

$$\begin{aligned} \sigma^2 &\in \{1/2, 1, 3/2, 2, 5/2\} \\ \beta &\in \{1/2, 3/4, 1, 4/3, 2/1\} \\ K &\in \{20, 40, 60, 80, 100\}, \end{aligned}$$

where $\beta$ varies the bias vector by $\boldsymbol{b} = [1/2, 3/4, 1, 4/3, 2/1]^T\beta$. Forty synthetic datasets are generated for each combination of $\sigma^2$, $\beta$, and $K$ values. The SAC-procedure runs for 200 iterations of which the first 100 are used for burn-in.

SAC under the Brier ($\text{SAC}_{\text{BRI}}$) and logarithm score ($\text{SAC}_{\text{LOG}}$) are compared with the *Exponentially Weighted Moving Average* (EWMA). EWMA, which serves as a baseline, can be understood by first denoting the (expertise-weighted) average forecast at time $t$ for the $k$th question with

$$\bar{p}_{t,k} = \sum_{j=1}^{J} \omega_j \left( \frac{1}{|E_j|} \sum_{i \in E_j} p_{i,t,k} \right) \tag{3.4}$$

where $E_j$ refers to an index set of all experts in the $j$th expertise group, and $\omega_j$ denotes the weight associated with the $j$th expertise group. The EWMA forecasts for the $k$th problem

Table 3.3: Summary measures of the estimation accuracy under synthetic data. As EWMA does not produce an estimate of the bias vector, its accuracy on the bias term cannot be reported.

| Hidden Process | | |
|---|---|---|
| Model | Quadratic Loss | Absolute Loss |
| $SAC_{BRI}$ | 0.00226 | 0.0334 |
| $SAC_{LOG}$ | 0.00200 | 0.0313 |
| EWMA | 0.00225 | 0.0339 |
| **Bias Vector** | | |
| Model | Quadratic Loss | Absolute Loss |
| $SAC_{BRI}$ | 0.147 | 0.217 |
| $SAC_{LOG}$ | 0.077 | 0.171 |

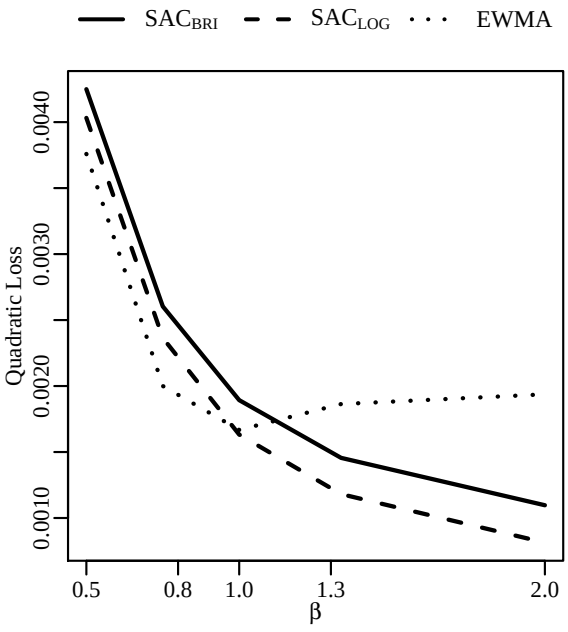



Figure 3.2: The marginal effect of $\beta$ on the average quadratic loss.

are then constructed recursively from

$$\hat{p}_{t,k}(\alpha) = \begin{cases} \bar{p}_{1,k}, & \text{for } t = 1, \\ \alpha \bar{p}_{t,k} + (1-\alpha)\hat{p}_{t-1,k}(\alpha), & \text{for } t > 1, \end{cases}$$

where $\alpha$ and $\boldsymbol{\omega}$ are learned from the training set by

$$(\hat{\alpha}, \hat{\boldsymbol{\omega}}) \quad = \quad \underset{\alpha, \omega_j \in [0,1]}{\arg\min} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \left(Z_k - \hat{p}_{t,k}(\alpha, \boldsymbol{\omega})\right)^2 \qquad \text{s.t.} \quad \sum_{j=1}^{J} \omega_j = 1$$

If $p_{t,k} = \text{logit}^{-1}(X_{t,k})$ and $\hat{p}_{t,k}$ is the corresponding probability estimated by the model, the model's accuracy to estimate the hidden process is measured with the quadratic loss, $(p_{t,k} - \hat{p}_{t,k})^2$, and the absolute loss, $|p_{t,k} - \hat{p}_{t,k}|$. Table 3.3 reports these losses averaged

over all conditions, simulations, and time points. The three competing methods, $\text{SAC}_{\text{BRI}}$, $\text{SAC}_{\text{LOG}}$, and EWMA, estimate the hidden process with great accuracy. Based on other performance measures that are not shown for the sake of brevity, all three methods suffer from an increasing level of noise in the expert logit-probabilities but can make efficient use of extra data.

Some interesting differences emerge from Figure 3.2 which shows the marginal effect of $\beta$ on the average quadratic loss. As can be expected, EWMA performs well when the experts are, on average, close to unbiased. Interestingly, SAC estimates the hidden process more accurately when the experts are over-confident (large $\beta$) compared to under-confident (small $\beta$). To understand this result, assume that the experts in the third group are highly under-confident. Their logit-probabilities are then expected to be closer to zero than the corresponding hidden states. After adding white noise to these expected logit-probabilities, they are likely to cross to the other side of zero. If the sampling step fixes $b_3 = 1$, as it does in our case, the third group is treated as unbiased and some of the constrained estimates of the hidden states are likely to be on the other side of zero as well. Unfortunately, this discrepancy cannot be corrected by the calibration step that is restricted to shifting the constrained estimates either closer or further away from zero but not across it. To maximize the likelihood of having all the constrained estimates on the right side of zero and hence avoiding the discrepancy, the reference point in the sampling step should be chosen with care. A helpful guideline is to fix the element of $\boldsymbol{b}$ that is *a priori* believed to be the largest.

The accuracy of the estimated bias vector is measured with the quadratic loss, $(b_j - \hat{b}_j)^2$, and the absolute loss, $|b_j - \hat{b}_j|$. Table 3.3 reports these losses averaged over all conditions, simulations, and elements of the bias vector. Unfortunately, EWMA does not produce an estimate of the bias vector. Therefore it cannot be used as a baseline for the estimation accuracy in this case. Given that the losses for $\text{SAC}_{\text{BRI}}$ and $\text{SAC}_{\text{LOG}}$ are quite small, they

estimate the bias vector accurately.

## 3.6 Geopolitical Data Results

This section presents results for the real-world data described in Section 3.2. The goal is to provide application specific insight by discussing the specific research objectives itemized in Section 3.1. First, however, we discuss two practical matters that must be taken into account when aggregating real-world probability forecasts.

### 3.6.1 Incoherent and Imbalanced Data

The first matter regards human experts making probability forecasts of 0.0 or 1.0 even if they are not completely sure of the outcome of the event. For instance, all 166 questions in our dataset contain both a zero and a one. Transforming such forecasts into the logit-space yields infinities that can cause problems in model estimation. To avoid this, Ariely et al. (2000) suggest changing $p = 0.00$ and $1.00$ to $p = 0.02$ and $0.98$, respectively. This is similar to *winsorising* that sets the extreme probabilities to a specified percentile of the data (see, e.g., Hastings et al. (1947) for more details on winsorising). Allard et al. (2012), on the other hand, consider only probabilities that fall within a constrained interval, say $[0.001, 0.999]$, and discard the rest. Given that this implies ignoring a portion of the data, we adopt a censoring approach similar to Ariely et al. (2000) by changing $p = 0.00$ and $1.00$ to $p = 0.01$ and $0.99$, respectively. Our results remain insensitive to the exact choice of censoring as long as this is done in a reasonable manner to keep the extreme probabilities from becoming highly influential in the logit-space.

The second matter is related to the distribution of the class labels in the data. If the set of occurrences is much larger than the set of non-occurrences (or *vice versa*), the dataset is called *imbalanced*. On such data the modeling procedure can end up over-focusing on the

larger class, and as a result, give very accurate forecast performance over the larger class at the cost of performing poorly over the smaller class (see, e.g., Chen (2009); Wallace and Dahabreh (2012)). Fortunately, it is often possible to use a well-balanced version of the data. The first step is to find a partition $S_0$ and $S_1$ of the question indices $\{1, 2, \ldots, K\}$ such that the equality $\sum_{k \in S_0} T_k = \sum_{k \in S_1} T_k$ is as closely approximated as possible. This is equivalent to an NP-hard problem known in computer science as the *Partition Problem*: determine whether a given set of positive integers can be partitioned into two sets such that the sums of the two sets equal to each other (see, e.g., Karmarkar and Karp (1982); Hayes (2002)). A simple solution is to use a greedy algorithm that iterates through the values of $T_k$ in descending order, assigning each $T_k$ to the subset that currently has the smaller sum (see, e.g., Kellerer et al. (2004); Gent and Walsh (1996) for more details on the *Partition Problem*). After finding a well-balanced partition, the next step is to assign the class labels such that the labels for the questions in $S_x$ are equal to $x$ for $x = 0$ or 1. Recall from Section 3.4.2 that $Z_k$ represents the event indicator for the $k$th question. To define a balanced set of indicators $\tilde{Z}_k$ for all $k \in S_x$, let

$$\tilde{Z}_k = x$$

$$\tilde{p}_{i,t,k} = \begin{cases} 1 - p_{i,t,k}, & \text{if } Z_k = 1 - x, \\ p_{i,t,k}, & \text{if } Z_k = x, \end{cases}$$

where $i = 1, \ldots, I_k$, and $t = 1, \ldots, T_k$. The resulting set

$$\left\{ \left( \tilde{Z}_k, \{\tilde{p}_{i,t,k} | i = 1, \ldots, I_k, t = 1, \ldots, T_k\} \right) \right\}_{k=1}^{K}$$

is a balanced version of the data. This procedure was used to balance our real-world dataset both in terms of events and time points. The final output splits the events exactly in half ($|S_0| = |S_1| = 83$) such that number of time points in the first and second halves are 8,737

and 8,738, respectively.

### 3.6.2 Out-of-Sample Aggregation

The goal of this section is to evaluate the accuracy of the aggregate probabilities made by SAC and several other procedures. The models are allowed to utilize a training set before making aggregations on an independent testing set. To clarify some of the upcoming notation, let $S_{train}$ and $S_{test}$ be index sets that partition the data into training and testing sets of sizes $|S_{train}| = N_{train}$ and $|S_{test}| = 166 - N_{train}$, respectively. This means that the $k$th question is in the training set if and only if $k \in S_{train}$. Before introducing the competing models, note that all choices of thinning and burn-in made in this section are conservative and have been made based on pilot runs of the models. This was done to ensure a posterior sample that has low autocorrelation and arises from a converged chain. The competing models are as follows.

1. *Simple Dynamic Linear Model (SDLM)*. This is equivalent to the dynamic model from Section 3.3 but with $\boldsymbol{b} = \mathbf{1}$ and $\beta = 1$. Thus,

$$\begin{aligned} \boldsymbol{Y}_{t,k} &= X_{t,k} + \boldsymbol{v}_{t,k} \\ X_{t,k} &= \gamma_k X_{t-1,k} + w_{t,k}, \end{aligned}$$

   where $X_{t,k}$ is the aggregate logit-probability. Given that this model does not share any parameters across questions, estimates of the hidden process can be obtained directly for the questions in the testing set without fitting the model first on the training set. The Gibbs sampler is run for 500 iterations of which the first 200 are used for burn-in. The remaining 300 iterations are thinned by discarding every other observation, leaving a final posterior sample of 150 observations. The average of this sample gives the final estimates.

2. *The Sample-And-Calibrate procedure both under the Brier ($SAC_{BRI}$) and the logarithmic score ($SAC_{LOG}$).* The model is first fit on the training set by running the sampling step for 3,000 iterations of which the first 500 iterations are used for burn-in. The remaining 2,500 observations are thinned by keeping every fifth observation. The calibration step is performed for the final 500 observations. The out-of-sample aggregation is done by running the sampling step for 500 iterations with each consecutive iteration reading in and conditioning on the next value of $\beta$ and $\boldsymbol{b}$ found during the training period. The first 200 iterations are used for burn-in. The remaining 300 iterations are thinned by discarding every other observation, leaving a final posterior sample of 150 observations. The average of this sample gives the final estimates.

3. *A fully Bayesian version of $SAC_{LOG}$ ($BSAC_{LOG}$).* Denote the calibrated logit probabilities and event indicators across all $K$ questions with $\boldsymbol{X}(1)$ and $\boldsymbol{Z}$, respectively. The posterior distribution of $\beta$ conditional on $\boldsymbol{X}(1)$ is given by

$$p(\beta|\boldsymbol{X}(1), \boldsymbol{Z}) \propto p(\boldsymbol{Z}|\beta, \boldsymbol{X}(1))p(\beta|\boldsymbol{X}(1)).$$

The likelihood is

$$p(\boldsymbol{Z}|\beta, \boldsymbol{X}(1)) \propto \prod_{k=1}^{K}\prod_{t=1}^{T_k} \quad \text{logit}^{-1}\left(X_{t,k}(1)/\beta\right)^{Z_k} \times \left(1 - \text{logit}^{-1}\left(X_{t,k}(1)/\beta\right)\right)^{1-Z_k} \tag{3.5}$$

As in Gelman et al. (2003), the prior for $\beta$ is chosen to be locally uniform, $p(1/\beta) \propto 1$. Given that this model estimates $X_{t,k}(1)$ and $\beta$ simultaneously, it is a little more flexible than SAC. Posterior estimates of $\beta$ can be sampled from (3.5) using generic sampling algorithms such as the Metropolis algorithm (Metropolis et al. (1953)) or slice sampling (Neal (2003)). Given that the sampling procedure conditions on the

event indicators, the full conditional distribution of the hidden states is not in a standard form. Therefore the Metropolis algorithm is also used for sampling the hidden states. Estimation is made with the same choices of thinning and burn-in as described under *Sample-And-Calibrate*.

4. Due to the lack of previous literature on dynamic aggregation of expert probability forecasts, the main competitors are exponentially weighted versions of procedures that have been proposed for static probability aggregation:

   (a) *Exponentially Weighted Moving Average (EWMA)* as described in Section 3.5.

   (b) *Exponentially Weighted Moving Logit Aggregator (EWMLA)*. This is a moving version of the aggregator $\hat{p}_G(\boldsymbol{b})$ that was introduced in Satopää et al. (2014). The EWMLA aggregate probabilities are found recursively from

$$\hat{p}_{t,k}(\alpha, \boldsymbol{b}) = \begin{cases} G_{1,k}(\boldsymbol{b}), & \text{for } t = 1, \\ \alpha G_{t,k}(\boldsymbol{b}) + (1-\alpha)\hat{p}_{t-1,k}(\alpha, \boldsymbol{b}), & \text{for } t > 1, \end{cases}$$

   where the vector $\boldsymbol{b} \in \mathbb{R}^J$ collects the bias terms of the expertise groups, and

$$G_{t,k}(\nu) = \left( \prod_{i=1}^{N_{t,k}} \left( \frac{p_{i,t,k}}{1-p_{i,t,k}} \right)^{\frac{b_{j(i,k)}}{N_{t,k}}} \right) \Bigg/ \left( 1 + \prod_{i=1}^{N_{t,k}} \left( \frac{p_{i,t,k}}{1-p_{i,t,k}} \right)^{\frac{b_{j(i,k)}}{N_{t,k}}} \right)$$

   The parameters $\alpha$ and $\boldsymbol{b}$ are learned from the training set by

$$(\hat{\alpha}, \hat{\boldsymbol{b}}) \quad = \quad \underset{\boldsymbol{b} \in \mathbb{R}^5, \alpha \in [0,1]}{\arg\min} \sum_{k \in S_{train}} \sum_{t=1}^{T_k} \left( Z_k - \hat{p}_{t,k}(\alpha, \boldsymbol{b}) \right)^2$$

   (c) *Exponentially Weighted Moving Beta-transformed Aggregator (EWMBA)*. The static version of the Beta-transformed aggregator was introduced in Ranjan and Gneiting (2010). A dynamic version can be obtained by replacing $G_{t,k}(\nu)$ in the

EWMLA description with $H_{\nu,\tau}(\bar{p}_{t,k})$, where $H_{\nu,\tau}$ is the cumulative distribution function of the Beta distribution and $\bar{p}_{t,k}$ is given by 3.4. The parameters $\alpha, \nu, \tau$, and $\boldsymbol{\omega}$ are learned from the training set by

$$
\begin{aligned}
(\hat{\alpha}, \hat{\nu}, \hat{\tau}, \hat{\boldsymbol{\omega}}) &= \underset{\nu,\tau>0\ \alpha,\omega_j\in[0,1]}{\arg\min} \sum_{k\in S_{train}} \sum_{t=1}^{T_k} (Z_k - \hat{p}_{t,k}(\alpha,\nu,\tau,\boldsymbol{\omega}))^2 \\
&\text{s.t.} \sum_{j=1}^{J} \omega_j = 1
\end{aligned}
$$

The competing models are evaluated via a 10-fold cross-validation[3] that first partitions the 166 questions into 10 sets such that each set has approximately the same number of questions (16 or 17 questions in our case) and the same number of time points (between 1,760 and 1,764 time points in our case). The evaluation then iterates 10 times, each time using one of the 10 sets as the testing set and the remaining 9 sets as the training set. Therefore each question is used nine times for training and exactly once for testing. The testing proceeds sequentially one testing question at a time as follows: First, for a question with a time horizon of $T_k$, give an aggregate probability at time $t = 2$ based on the first two days. Compute the Brier score for this probability. Next give an aggregate probability at time $t = 3$ based on the first three days and compute the Brier score for this probability. Repeat this process for all of the $T_k - 1$ days. This leads to $T_k - 1$ Brier scores per testing question and a total of 17,475 Brier scores across the entire dataset.

Table 3.4 summarizes these scores in different ways. The first option, denoted by *Scores by Day*, weighs each question by the number of days the question remained open. This is performed by computing the average of the 17,475 scores. The second option, denoted by *Scores by Problem*, gives each question an equal weight regardless how long the question remained open. This is done by first averaging the scores within a question and then aver-

[3] A 5-fold cross-validation was also performed. The results were, however, very similar to the 10-fold cross-validation and hence not presented in the paper.

Table 3.4: Brier Scores based on 10-fold cross-validation. *Scores by Day* weighs a question by the number of days the question remained open. *Scores by Problem* gives each question an equal weight regardless how long the question remained open. The bolded values indicate the best scores in each column. The values in the parenthesis represent standard errors in the scores.

| Scores by Day | | | | |
|---|---|---|---|---|
| Model | All | Short | Medium | Long |
| SDLM | 0.100 (0.156) | 0.066 (0.116) | 0.098 (0.154) | 0.102 (0.157) |
| $BSAC_{LOG}$ | 0.097 (0.213) | **0.053** (0.147) | 0.100 (0.215) | 0.098 (0.215) |
| $SAC_{BRI}$ | 0.096 (0.190) | 0.056 (0.134) | 0.097 (0.190) | 0.098 (0.192) |
| $SAC_{LOG}$ | **0.096** (0.191) | 0.056 (0.134) | **0.096** (0.189) | **0.098** (0.193) |
| EWMBA | 0.104 (0.204) | 0.057 (0.120) | 0.113 (0.205) | 0.105 (0.206) |
| EWMLA | 0.102 (0.199) | 0.061 (0.130) | 0.111 (0.214) | 0.103 (0.200) |
| EWMA | 0.111 (0.146) | 0.080 (0.101) | 0.116 (0.152) | 0.112 (0.146) |
| Scores by Problem | | | | |
| Model | All | Short | Medium | Long |
| SDLM | 0.089 (0.116) | 0.064 (0.085) | 0.106 (0.141) | 0.092 (0.117) |
| $BSAC_{LOG}$ | 0.083 (0.160) | **0.052** (0.103) | 0.110 (0.198) | 0.085 (0.162) |
| $SAC_{BRI}$ | 0.083 (0.142) | 0.055 (0.096) | 0.106 (0.174) | 0.085 (0.144) |
| $SAC_{LOG}$ | **0.082** (0.142) | 0.055 (0.096) | **0.105** (0.174) | **0.085** (0.144) |
| EWMBA | 0.091 (0.157) | 0.057 (0.095) | 0.121 (0.187) | 0.093 (0.164) |
| EWMLA | 0.090 (0.159) | 0.064 (0.109) | 0.120 (0.200) | 0.090 (0.159) |
| EWMA | 0.102 (0.108) | 0.080 (0.075) | 0.123 (0.130) | 0.103 (0.110) |

aging the average scores across all the questions. Both scores can be further broken down into subcategories by considering the length of the questions. The final three columns of Table 3.4 divide the questions into *Short* questions (30 days or fewer), *Medium* questions (between 31 and 59 days), and *Long* Problems (60 days or more). The number of questions in these subcategories were 36, 32 and 98, respectively. The bolded scores indicate the best score in each column. The values in the parenthesis quantify the variability in the scores: Under *Scores by Day* the values give the standard errors of all the scores. Under *Scores by Problem*, on the other hand, the values represent the standard errors of the average scores of the different questions.

As can be seen in Table 3.4, $\text{SAC}_{\text{LOG}}$ achieves the lowest score across all columns except *Short* where it is outperformed by $\text{BSAC}_{\text{LOG}}$. It turns out that $\text{BSAC}_{\text{LOG}}$ is over-confident (see Section 3.6.3). This means that $\text{BSAC}_{\text{LOG}}$ underestimates the uncertainty in the events and outputs aggregate probabilities that are typically too near 0.0 or 1.0. This results into highly variable performance. The short questions generally involved very little uncertainty. On such easy questions, overconfidence can pay off frequently enough to compensate for a few large losses arising from the overconfident and drastically incorrect forecasts.

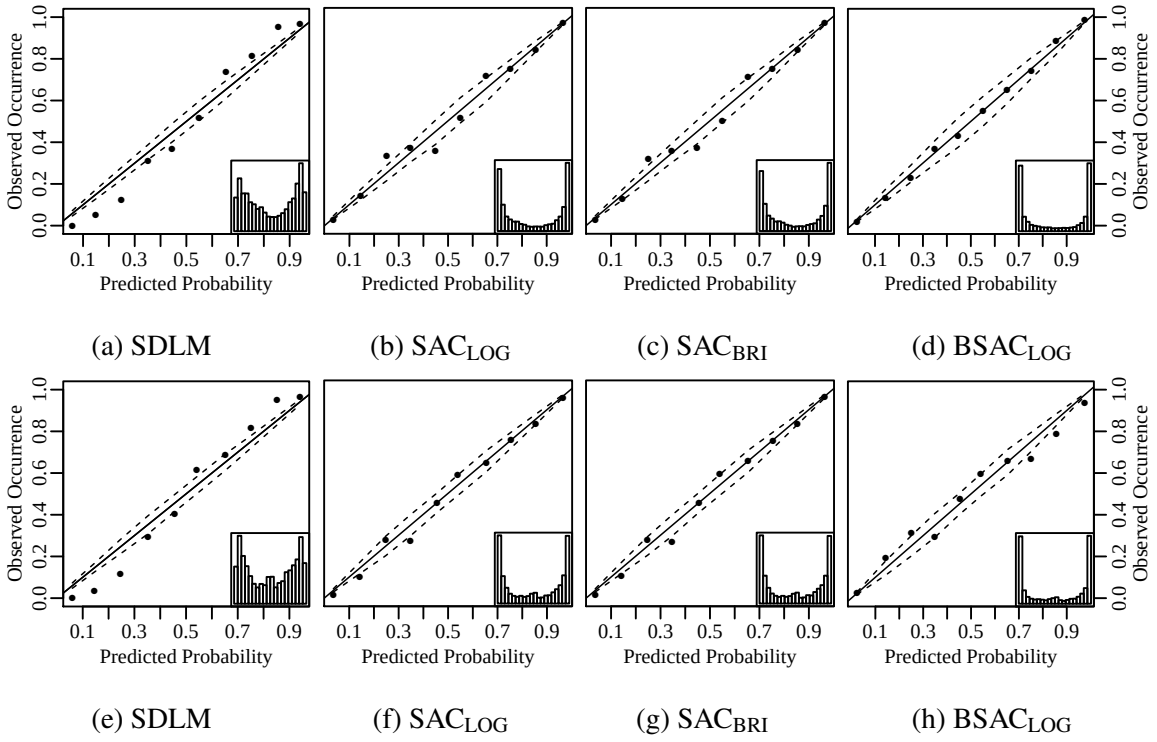
SDLM, on the other hand, lacks sharpness and is highly under-confident (see Section 3.6.3). This behavior is expected as the experts are under-confident at the group-level (see Section 3.6.4) and SDLM does not use the training set to explicitly calibrate its aggregate probabilities. Instead, it merely smooths the forecasts given by the experts. The resulting aggregate probabilities are therefore necessarily conservative, resulting into high average scores with low variability.

Similar behavior is exhibited by EWMA that performs the worst of all the competing models. The other two exponentially weighted aggregators, EWMLA and EWMBA, make efficient use of the training set and present moderate forecasting performance in most columns of Table 3.4. Neither approach, however, appears to dominate the other. The high variability and average of their performance scores indicate that their performance suffers from over-confidence.

### 3.6.3 In- and Out-of-Sample Sharpness and Calibration

A calibration plot is a simple tool for visually assessing the sharpness and calibration of a model. The idea is to plot the aggregate probabilities against the observed empirical frequencies. Therefore any deviation from the diagonal line suggests poor calibration. A model is considered under-confident (or over-confident) if the points follow an S-shaped

(or ƨ-shaped) trend. To assess sharpness of the model, it is common practice to place a histogram of the given forecasts in the corner of the plot. Given that the data were balanced, any deviation from the the baseline probability of 0.5 suggests improved sharpness.




(a) SDLM (b) $SAC_{LOG}$ (c) $SAC_{BRI}$ (d) $BSAC_{LOG}$

(e) SDLM (f) $SAC_{LOG}$ (g) $SAC_{BRI}$ (h) $BSAC_{LOG}$

Figure 3.3: The top and bottom rows show in- and out-of-sample calibration and sharpness, respectively.

The top and bottom rows of Figure 3.3 present calibration plots for SDLM, $SAC_{LOG}$, $SAC_{BRI}$, and $BSAC_{LOG}$ under in- and out-of-sample probability aggregation, respectively. Each setting is of interest in its own right: Good in-sample calibration is crucial for model interpretability. In particular, if the estimated crowd belief is well-calibrated, then the elements of the bias vector $\boldsymbol{b}$ can be used to study the amount of under- or over-confidence in the different expertise groups. Good out-of-sample calibration and sharpness, on the other hand, are necessary properties in decision making. To guide our assessment, the dashed bands around the diagonal connect the point-wise, Bonferroni-corrected (Bonfer-

roni (1936)) 95% lower and upper critical values under the null hypothesis of calibration. These have been computed by running the bootstrap technique described in Bröcker and Smith (2007) for 10,000 iterations. The in-sample predictions were obtained by running the models for 10,200 iterations, leading to a final posterior sample of 1,000 observations after thinning and using the first 200 iterations for burn-in. The out-of-sample predictions were given by the 10-fold cross-validation discussed in Section 3.6.2.

Overall, SAC is sharp and well-calibrated both in- and out-of-sample with only a few points barely falling outside the *point-wise* critical values. Given that the calibration does not change drastically from the top to the bottom row, SAC can be considered robust against over-fitting. This, however, is not the case with $\text{BSAC}_{\text{LOG}}$ that is well-calibrated in-sample but presents over-confidence out-of-sample. Figures 3.3a and 3.3e serve as baselines by showing the calibration plots for SDLM. Given that this model does not perform any explicit calibration, it is not surprising to see most points outside the critical values. The pattern in the deviations suggests strong under-confidence. Furthermore, the inset histogram reveals drastic lack of sharpness. Therefore SAC can be viewed as a well-performing compromise between SDLM and $\text{BSAC}_{\text{LOG}}$ that avoids over-confidence without being too conservative.

### 3.6.4 Group-Level Expertise Bias

This section explores the bias among the five expertise groups in our dataset. Figure 3.4 compares the posterior distributions of the individual elements of $\boldsymbol{b}$ with side-by-side box-plots. Given that the distributions fall completely below the *no-bias* reference-line at 1.0, all the expertise groups are deemed under-confident. Even though the exact level of under-confidence is affected slightly by the extent to which the extreme probabilities are censored (see Section 3.6.1), the qualitative results in this section remain insensitive to different levels of censoring.

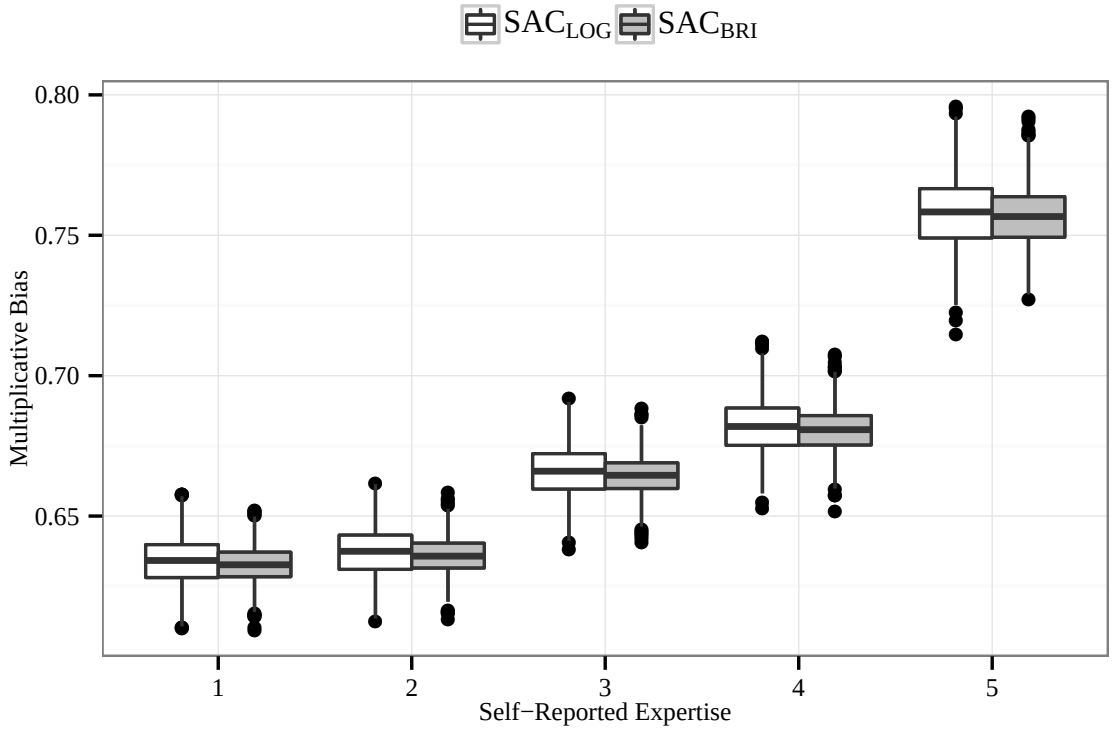



Figure 3.4: Posterior distributions of $b_j$ for $j = 1, \ldots, 5$.

Figure 3.4 shows that under-confidence decreases as expertise increases. The posterior probability that the most expert group is the least under-confident is approximately equal to $1.0$, and the posterior probability of a strictly decreasing level of under-confidence is approximately 0.87. The latter probability is driven down by the inseparability of the two groups with the lowest levels of self-reported expertise. This inseparability suggests that the experts are poor at assessing how little they know about a topic that is strange to them. If these groups are combined into a single group, the posterior probability of a strictly decreasing level of under-confidence is approximately 1.0.

The decreasing trend in under-confidence can be viewed as a process of Bayesian updating. A completely ignorant expert aiming to minimize a reasonable loss function, such as the Brier score, has no reason to give anything but 0.5 as his probability forecast. However, as soon as the expert gains some knowledge about the event, he produces an updated

forecast that is a compromise between his initial forecast and the new information acquired. The updated forecast is therefore conservative and too close to 0.5 as long as the expert remains only partially informed about the event. If most experts fall somewhere on this spectrum between ignorance and full information, their average forecast tends to fall strictly between 0.5 and the most-informed probability forecast (see Baron et al. (2014) for more details). Given that expertise is to a large extent determined by subject-matter knowledge, the level of under-confidence can be expected to decrease as a function of the group's level of self-reported expertise.

Finding under-confidence in all the groups may seem like a surprising result given that many previous studies have shown that experts are often over-confident (see, e.g., Lichtenstein et al. (1977); Morgan (1992); Bier (2004) for a summary of numerous calibration studies). It is, however, worth emphasizing three points: First, our result is a statement about groups of experts and hence does not invalidate the possibility of the individual experts being overconfident. To make conclusions at the individual-level based on the group-level bias terms would be considered an *ecological inference fallacy* (see, e.g., Lubinski and Humphreys (1996)). Second, the experts involved in our dataset are overall very well calibrated (Mellers et al. (2014)). A group of well-calibrated experts, however, can produce an aggregate forecast that is under-confident. In fact, if the aggregate is linear, the group is necessarily under-confident (see Theorem 1 of Ranjan and Gneiting (2010)). Third, according to Erev et al. (1994) the level of confidence depends on the way the data were analyzed. They explain that experts' probability forecasts suggest under-confidence when the forecasts are averaged or presented as a function of independently defined objective probabilities, i.e. the probabilities given by $\mathrm{logit}^{-1}(X_{t,k})$ in our case. This is similar to our context and opposite to many empirical studies on confidence calibration.

### 3.6.5 Question Difficulty and Other Measures

One advantage of our model arises from its ability to produce estimates of interpretable question-specific parameters $\gamma_k$, $\sigma_k^2$, and $\tau_k^2$. These quantities can be combined in many interesting ways to answer questions about different groups of experts or the questions themselves. For instance, being able to assess the difficulty of a question could lead to more principled ways of aggregating performance measures across questions or to novel insight on the kind of questions that are found difficult by experts (see, e.g., a discussion on the *Hard-Easy Effect* in Wilson (1994)). To illustrate, recall that higher values of $\sigma_k^2$ suggest greater disagreement among the participating experts. Given that experts are more likely to disagree over a difficult question than an easy one, it is reasonable to assume that $\sigma_k^2$ has a positive relationship with question difficulty. An alternative measure is given by $\tau_k^2$ that quantifies the volatility of the underlying circumstances that ultimately decide the outcome of the event. Therefore a high value of $\tau_k^2$ can cause the outcome of the event to appear unstable and difficult to predict.

As a final illustration of our model, we return to the two example questions introduced Figure 3.1. Given that $\hat{\sigma}_k^2 = 2.43$ and $\hat{\sigma}_k^2 = 1.77$ for the questions depicted in Figures 3.1a and 3.1b, respectively, the first question provokes more disagreement among the experts than the second one. Intuitively this makes sense because the target event in Figure 3.1a is determined by several conditions that may change radically from one day to the next while the target event in Figure 3.1b is determined by a relatively steady stock market index. Therefore it is not surprising to find that in Figure 3.1a $\hat{\tau}_k^2 = 0.269$, which is much higher than $\hat{\tau}_k^2 = 0.039$ in Figure 3.1b. We may conclude that the first question is inherently more difficult than the second one.

## 3.7 Discussion

This paper began by introducing a rather unorthodox but nonetheless realistic time-series setting where probability forecasts are made very infrequently by a heterogeneous group of experts. The resulting data is too sparse to be modeled well with standard time-series methods. In response to this lack of appropriate modeling procedures, we propose an interpretable time-series model that incorporates self-reported expertise to capture a sharp and well-calibrated estimate of the crowd belief. This procedure extends the forecasting literature into an under-explored area of probability aggregation.

Our model preserves parsimony while addressing the main challenges in modeling sparse probability forecasting data. Therefore it can be viewed as a basis for many future extensions. To give some ideas, recall that most of the model parameters were assumed constant over time. It is intuitively reasonable, however, that these parameters behave differently during different time intervals of the question. For instance, the level of disagreement (represented by $\sigma_k^2$ in our model) among the experts can be expected to decrease towards the final time point when the question resolves. This hypothesis could be explored by letting $\sigma_{t,k}^2$ evolve dynamically as a function of the previous term $\sigma_{t-1,k}^2$ and random noise.

This paper modeled the bias separately within each expertise group. This is by no means restricted to the study of bias or its relation to self-reported expertise. Different parameter dependencies could be constructed based on many other expert characteristics, such as gender, education, or specialty, to produce a range of novel insights on the forecasting behavior of experts. It would also be useful to know how expert characteristics interact with question types, such as economic, domestic, or international. The results would be of interest to the decision-maker who could use the information as a basis for hiring only a high-performing subset of the available experts.

Other future directions could remove some of the obvious limitations of our model. For

instance, recall that the random components are assumed to follow a normal distribution. This is a strong assumption that may not always be justified. Logit-probabilities, however, have been modeled with the normal distribution before (see, e.g., Erev et al. (1994)). Furthermore, the normal distribution is a rather standard assumption in psychological models (see, e.g., signal-detection theory in Tanner Jr and Swets (1954)).

A second limitation resides in the assumption that both the observed and hidden processes are expected to grow linearly. This assumption could be relaxed, for instance, by adding higher order terms to the model. A more complex model, however, is likely to sacrifice interpretability. Given that our model can detect very intricate patterns in the crowd belief (see Figure 3.1), compromising interpretability for the sake of facilitating non-linear growth is hardly necessary.

A third limitation appears in an online setting where new forecasts are received at a fast rate. Given that our model is fit in a retrospective fashion, it is necessary to refit the model every time a new forecast becomes available. Therefore our model can be applied only to offline aggregation and online problems that tolerate some delay. A more scalable and efficient alternative would be to develop an aggregator that operates recursively on streams of forecasts. Such a *filtering* perspective would offer an aggregator that estimates the current crowd belief accurately without having to refit the entire model each time a new forecast arrives. Unfortunately, this typically implies being less accurate in estimating the model parameters such as the bias term. However, as estimation of the model parameters was addressed in this paper, designing a filter for probability forecasts seems like the next natural development in time-series probability aggregation.

## 3.8 Acknowledgements

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Ac-

tivity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. We deeply appreciate the project management skills and work of Terry Murray and David Wayrynen, which went far beyond the call-of-duty on this project.

# 4

# Modeling Probability Forecasts via Information Diversity*


## Abstract

Randomness in scientific estimation is generally assumed to arise from unmeasured or uncontrolled factors. However, when combining subjective probability estimates, heterogeneity stemming from people's cognitive or information diversity is often more important than measurement noise. This paper presents a novel framework that uses partially overlapping information sources. A specific model is proposed within that framework and applied to the task of aggregating the probabilities given by a group of forecasters who predict whether an event will occur or not. Our model describes the distribution of information across forecasters in terms of easily interpretable parameters and shows how the optimal amount of *extremizing* of the average probability forecast (shifting it closer to its nearest extreme) varies as a function of the forecasters' information overlap. Our model thus gives a more principled understanding of the historically *ad hoc* practice of extremizing average forecasts. Supplementary material for this article is available online.


*Joint work with Robin Pemantle and Lyle H. Ungar

## 4.1 Introduction and Overview

### 4.1.1 The Forecast Aggregation Problem

Probability forecasting is the science of giving probability estimates for future events. Typically more than one different forecast is available on the same event. Instead of trying to guess which prediction is the most accurate, the predictions should be combined into a single consensus forecast (Armstrong, 2001). Unfortunately, the forecasts can be combined in many different ways, and the choice of the combination rule can largely determine the predictive quality of the final aggregate. This is the principal motivation for the problem of *forecast aggregation* that aims to combine multiple forecasts into a single forecast with optimal properties.

There are two general approaches to forecast aggregation: empirical and theoretical. Given a training set with multiple forecasts on events with known outcomes, the empirical approach experiments with different aggregation techniques and chooses the one that yields the best performance on the training set. The theoretical approach, on the other hand, first constructs a probability model and then computes the optimal aggregation procedure under the model assumptions. Both approaches are important. Theory-based procedures that do not perform well in practice are ultimately of limited use. On the other hand, an empirical approach without theoretical underpinnings lacks both credibility (why should we believe it?) and guidance (in which direction can we look for improvement?). As will be discussed below, the history of forecast aggregation to date is largely empirical.

The main contribution of this paper is a plausible theoretical framework for forecast aggregation called the *partial information framework*. Under this framework, forecast heterogeneity stems from information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles on the presidential election may use distinct parts of the information and hence report different predictions of

a candidate winning. Second, the framework allows us to interpret existing aggregators and illuminate aspects that can be improved. This paper specifically aims to clarify the practice of *probability extremizing*, i.e., shifting an average aggregate closer to its nearest extreme. Extremizing is an empirical technique that has been widely used to improve the predictive performance of many simple aggregators such as the average probability. Lastly, the framework is applied to a specific model under which the optimal aggregator can be computed.

### 4.1.2 Bias, Noise, and Forecast Assessment

Consider an event $A$ and an indicator function $\mathbf{1}_A$ that equals one or zero depending whether $A$ happens or not, respectively. There are two common yet philosophically different approaches to linking $A$ with the probability forecasts. The first assumes $\mathbf{1}_A \sim \text{Bernoulli}(\theta)$, where $\theta$ is deemed a "true" or "objective" probability for $A$, and then treats a probability forecast $p$ as an estimator of $\theta$ (see, e.g., Lai et al. 2011, and Section 4.2.2 for further discussion). The second approach, on the other hand, treats $p$ as an estimator of $\mathbf{1}_A$. This links the observables directly and avoids the controversial concept of a "true" probability; for this reason it is the approach adopted in this paper.

As is the case with all estimators, the forecast's deviation from the truth can be broken into bias and noise. Given that these components are typically handled by different mechanisms, it is important, on the theoretical level, to consider them as two separate problems. This paper focuses on noise reduction. Therefore, each forecaster is considered *calibrated*. Here calibration is defined in terms of conditional expectation and hence represents a property of the underlying joint distribution of $\mathbf{1}_A$ and $p$. More specifically, the forecast $p$ is calibrated for the outcome $\mathbf{1}_A$ if $\mathbb{P}(\mathbf{1}_A = 1|p) = \mathbb{E}(\mathbf{1}_A|p) = p$ almost surely. This form of calibration was alluded to by Murphy and Winkler (1987b) and mentioned possibly even earlier than that. Over the years it has become common in the statistical and meteorological

forecasting literature (see, e.g., Ranjan and Gneiting 2010; Jolliffe and Stephenson 2012, Section 7.2.2. for recent references). It is, however, different from the notion of empirical calibration discussed by Dawid (1982), Foster and Vohra (1998), and many others.

A forecast (individual or aggregate) is typically assessed with a loss function $L(p, \mathbf{1}_A)$. A loss function is called *proper* or *revealing* if the Bayesian optimal strategy is to tell the truth. In other words, if the subjective probability estimate is $p$, then $t = p$ should minimize the expected loss $pL(t,1)+(1-p)L(t,0)$. Therefore, if a group of sophisticated forecasters operates under a proper loss function, the assumption of calibrated forecasts is, to some degree, self-fulfilling. There are, however, many different proper loss functions, and an estimator that outperforms another under one loss function will not necessarily do so under a different one. For example, minimizing the quadratic loss function $(p-\mathbf{1}_A)^2$, also known as the Brier score, gives the estimator with the least variance. This paper concentrates on minimizing the variance of the aggregators, though much of the discussion holds under general proper loss functions. See Hwang and Pemantle (1997) for a discussion of proper loss functions.

### 4.1.3 The Partial Information Framework

The construction of the partial information framework begins with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable event $A \in \mathcal{F}$ to be forecasted by $N$ forecasters. These forecasters operate under the same probability model but make predictions based on different information sets. More specifically, in any Bayesian setup, with a proper loss function, it is more or less tautological that Forecaster $i$ reports $p_i := \mathbb{E}(\mathbf{1}_A \,|\, \mathcal{F}_i)$, where $\mathcal{F}_i \subseteq \mathcal{F}$ is the *information set* used by the forecaster. Therefore $\mathcal{F}_i \neq \mathcal{F}_j$ if $p_i \neq p_j$, and forecast heterogeneity stems purely from *information diversity*. Note, however, that if Forecaster $i$ uses a simple rule, $\mathcal{F}_i$ may not be the full $\sigma$-field of information available to the forecaster but rather a smaller $\sigma$-field corresponding to the information used by the rule. For example,

when forecasting the re-election of the president, a forecaster obeying the dictum "it's the economy, stupid!" might utilize a $\sigma$-field containing only economic indicators. Furthermore, if two forecasters have access to the same $\sigma$-field, they may decide to use different sub-$\sigma$-fields, leading to different predictions. Therefore, information diversity does not only arise from differences in the available information, but also from how the forecasters decide to use it.

The person performing the aggregation is assumed to know only $\mathcal{F}_0 = \{\emptyset, \Omega\}$, namely the trivial $\sigma$-field. Given that every forecaster knows at least as much, the aggregator readily adopts any forecaster's prediction without modification. Therefore, each forecaster is considered to be an "expert" in the sense introduced in DeGroot (1988) and later discussed in Dawid et al. (1995).

Under the partial information framework the forecasts are calibrated. This can be verified by direct computation as follows:

$$\mathbb{E}(\mathbf{1}_A|p_i) = \mathbb{E}\{\mathbb{E}(\mathbf{1}_A|p_i, \mathcal{F}_i)|p_i\} = \mathbb{E}\{\mathbb{E}(\mathbf{1}_A|\mathcal{F}_i)|p_i\} = \mathbb{E}(p_i|p_i) = p_i.$$

Conversely, if $p_i$ is any calibrated forecast, then $p_i = \mathbb{E}(\mathbf{1}_A|\mathcal{G}_i)$, where $\mathcal{G}_i = \sigma(p_i) \subseteq \mathcal{F}_i$ is the $\sigma$-field generated by $p_i$. This shows constructively that assuming the general form $p_i = \mathbb{E}(\mathbf{1}_A \,|\, \mathcal{F}_i)$ does not pose any additional restrictions but arises directly from the assumption of calibration and the existence of an underlying probability model. The $\sigma$-field $\mathcal{G}_i$, however, corresponds to the information revealed by the forecast and hence may not be equal to the full $\sigma$-field of information actually used by the forecaster, namely $\mathcal{F}_i$.

The distinction between $\mathcal{F}_i$ and $\mathcal{G}_i$ introduces two benchmarks for aggregation efficiency. The first is the *oracular* aggregator $p' := \mathbb{E}(\mathbf{1}_A \,|\, \mathcal{F}')$, where $\mathcal{F}'$ is the $\sigma$-field generated by the union of the information sets $\{\mathcal{F}_i : i = 1, \ldots, N\}$. This field represents all the information used by the forecasters. Given that aggregation cannot be improved beyond using all the information of the forecasters, the oracular aggregator represents a theoretical

optimum and is therefore a reasonable upper bound on estimation efficiency.

In practice, however, information comes to the aggregator only through the forecasts $\{p_i : i = 1, \ldots, N\}$. Given that $\mathcal{F}'$ generally cannot be constructed from these forecasts alone, no practically feasible aggregator can be expected to perform as well as $p'$. Therefore, a more achievable benchmark is the *revealed* aggregator $p'' := \mathbb{E}(\mathbf{1}_A \mid \mathcal{F}'')$, where $\mathcal{F}''$ is the $\sigma$-field generated (or revealed) by the forecasts $\{p_i : i = 1, \ldots, N\}$, or equivalently by the union of the generated $\sigma$-fields $\{\mathcal{G}_i : i = 1, \ldots, N\}$.

Even though the partial information framework, as specified above, is too theoretical for direct application, it highlights the crucial components of information aggregation and hence facilitates formulation of more specific models within the framework. This paper develops such a model and calls it the Gaussian partial information model. Under this model, the information among the forecasters is summarized by a covariance structure. This provides sufficient flexibility to allow for construction of many application-specific aggregators.

### 4.1.4 Organization of the Paper

The next section reviews prior work on forecast aggregation and relates it to the partial information framework. Section 4.3 discusses illuminating examples and motivates the Gaussian partial information model. Section 4.4 compares the oracular aggregator with the average probit score, thereby explaining the empirical practice of probability extremizing. Section 4.5 derives the revealed aggregator and evaluates one of its sub-cases on real-world forecasting data. The final section concludes with a summary and discussion of future research.

## 4.2 Prior Work on Aggregation

### 4.2.1 The Interpreted Signal Framework

Hong and Page (2009) introduce the *interpreted signal framework* in which the forecaster's prediction is based on a personal interpretation of (a subset of) the factors or cues that influence the future event to be predicted. Differences among the predictions are ascribed to differing interpretation procedures. For example, if two forecasters follow the same political campaign speech, one forecaster may focus on the content of the speech while the other may concentrate largely on the audience interaction. Even though the forecasters receive the same information, they interpret it differently and therefore are likely to report different estimates of the probability that the candidate wins the election. Therefore forecast heterogeneity is assumed to stem from "cognitive diversity".

This is a very reasonable assumption that has been analyzed and utilized in many other settings. For example, Parunak et al. (2013) demonstrate that optimal aggregation of interpreted forecasts is not constrained to the convex hull of the forecasts; Broomell and Budescu (2009) analyze inter-forecaster correlation under the assumption that the cues can be mapped to the individual forecasts via different linear regression functions. To the best of our knowledge, no previous work has discussed a formal framework that explicitly links the interpreted forecasts to their target quantity. Consequently, the interpreted signal framework, as proposed, has remained relatively abstract. The partial information framework, however, formalizes the intuition behind it and permits models with quantitative predictions.

### 4.2.2 The Measurement Error Framework

In the absence of a quantitative interpreted signal model, prior applications have typically relied on the *measurement error framework* that generates forecast heterogeneity from a

probability distribution. More specifically, the framework assumes a "true" probability $\theta$, interpreted as the forecast made by an ideal forecaster, for the event $A$. The forecasters then "measure" some transformation of this probability $\phi(\theta)$ with mean-zero idiosyncratic error. Therefore each forecast is an independent draw from a common probability distribution centered at $\phi(\theta)$, and a recipe for an aggregate forecast is given by the average

$$\phi^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\phi(p_i)\right\}. \tag{4.1}$$

Common choices of $\phi(p)$ are the identity $\phi(p) = p$, the log-odds $\phi(p) = \log\{p/(1-p)\}$, and the probit $\phi(p) = \Phi^{-1}(p)$, giving three aggregators denoted in this paper with $\overline{p}$, $p_{\log}$, and $p_{\text{probit}}$, respectively. These *averaging aggregators* represents the main advantage of the measurement error framework: simplicity.

Unfortunately, there are a number of disadvantages. First, given that the averaging aggregators target $\phi(\theta)$ instead of $\mathbf{1}_A$, important properties such as calibration cannot be expected. In fact, the averaging aggregators are uncalibrated and under-confident, i.e., too close to $1/2$, even if the individual forecasts are calibrated (Ranjan and Gneiting, 2010).

Second, the underlying model is rather implausible. Relying on a true probability $\theta$ is vulnerable to many philosophical debates, and even if one eventually manages to convince one's self of the existence of such a quantity, it is difficult to believe that the forecasters are actually seeing $\phi(\theta)$ with independent noise. Therefore, whereas the interpreted signal framework proposes a micro-level explanation, the measurement error model does not; at best, it forces us to imagine that the forecasters are all in principle trying to apply the same procedures to the same data but are making numerous small mistakes.

Third, the averaging aggregators do not often perform very well in practice. For one thing, Hong and Page (2009) demonstrate that the standard assumption of conditional independence poses an unrealistic structure on interpreted forecasts. Any averaging aggregator is also constrained to the convex hull of the individual forecasts, which further contradicts

the interpreted signal framework (Parunak et al., 2013) and can be far from optimal on many datasets.

### 4.2.3 Empirical Approaches

If one is not concerned with theoretical justification, an obvious approach is to perturb one of these estimators and observe whether the adjusted estimator performs better on some data set of interest. Given that the measurement error framework produces under-confident aggregators, a popular adjustment is to *extremize*, that is, to shift the average aggregates closer to the nearest extreme (either zero or one). For instance, Ranjan and Gneiting (2010) extremize $\overline{p}$ with the CDF of a beta distribution; Satopää et al. (2014) use a logistic regression model to derive an aggregator that extremizes $p_{\log}$; Baron et al. (2014) give two intuitive justifications for extremizing and discuss an extremizing technique that has previously been used by a number of investigators (Erev et al. 1994; Shlomi and Wallsten 2010; and even Karmarkar 1978); Mellers et al. (2014) show empirically that extremizing can improve aggregate forecasts of international events.

These and many other studies represent the unwieldy position of the current state-of-the-art aggregators: they first compute an average based on a model that is likely to be at odds with the actual process of probability forecasting, and then aim to correct the induced bias via *ad hoc* extremizing techniques. Not only does this leave something to be desired from an explanatory point of view, these approaches are also subject to overfitting. Most importantly, these techniques provide little insight beyond the amount of extremizing itself and hence lack a clear direction of continued improvement. The present paper aims to remedy this situation by explaining extremization with the aid of a theoretically based estimator, namely the oracular aggregator.

## 4.3 The Gaussian Partial Information Model

### 4.3.1 Motivating Examples

A central component of the partial information models is the structure of the information overlap that is assumed to hold among the individual forecasters. It therefore behooves us to begin with some simple examples to show that the optimal aggregate is not well defined without assumptions on the information structure among the forecasters.

**Example 4.3.1.** Consider a basket containing a fair coin and a two-headed coin. Two forecasters are asked to predict whether a coin chosen at random is in fact two-headed. Before making their predictions, the forecasters observe the result of a single flip of the chosen coin. Suppose the flip comes out HEADS. Based on this observation, the correct Bayesian probability estimate is $2/3$. If both forecasters see the result of the same coin flip, the optimal aggregate is again $2/3$. On the other hand, if they observe different (conditionally independent) flips of the same coin, the optimal aggregate is $4/5$.

In this example, it is not possible to distinguish between the two different information structures simply based on the given predictions, and neither $2/3$ nor $4/5$ can be said to be a better choice for the aggregate forecast. Therefore, we conclude that it is necessary to incorporate an assumption as to the structure of the information overlap, and that the details must be informed by the particular instance of the problem. The next example shows that even if the forecasters observe marginally independent events, further details in the structure of information can still greatly affect the optimal aggregate forecast.

**Example 4.3.2.** Let $\Omega = \{A, B, C, D\} \times \{0, 1\}$ be a probability space with eight points. Consider a measure $\mu$ that assigns probabilities $\mu(A, 1) = a/4, \mu(A, 0) = (1 - a)/4$, $\mu(B, 1) = b/4, \mu(B, 0) = (1 - b)/4$, and so forth. Define two events

$$S_1 = \{(A, 0), (A, 1), (B, 0), (B, 1)\},$$

$$S_2 = \{(A,0),(A,1),(C,0),(C,1)\}.$$

Therefore, $S_1$ is the event that the first coordinate is $A$ or $B$, and $S_2$ is the event that the first coordinate is $A$ or $C$. Consider two forecasters and suppose Forecaster $i$ observes $S_i$. Therefore the $i$th Forecaster's information set is given by the $\sigma$-field $\mathcal{F}_i$ containing $S_i$ and its complement. Their $\sigma$-fields are independent. Now, let $G$ be the event that the second coordinate is 1. Forecaster 1 reports $p_1 = \mathbb{P}(G|\mathcal{F}_1) = (a+b)/2$ if $S_1$ occurs; otherwise, $p_1 = (c+d)/2$. Forecaster 2, on the other hand, reports $p_2 = \mathbb{P}(G|\mathcal{F}_2) = (a+c)/2$ if $S_2$ occurs; otherwise, $p_2 = (b+d)/2$. If $\varepsilon$ is added to $a$ and $d$ but subtracted from $b$ and $c$, the forecasts $p_1$ and $p_2$ do not change, nor does it change the fact that each of the four possible pairs of forecasts has probability $1/4$. Therefore all observables are invariant under this perturbation. If Forecasters 1 and 2 report $(a+b)/2$ and $(a+c)/2$, respectively, then the aggregator knows, by considering the intersection $S_1 \cap S_2$, that the first coordinate is $A$. Consequently, the optimal aggregate forecast is $a$, which is most definitely affected by the perturbation.

This example shows that the aggregation problem can be affected by the fine structure of information overlap. It is, however, unlikely that the structure can ever be known with the precision postulated in this simple example. Therefore it is necessary to make reasonable assumptions that yield plausible yet generic information structures.

### 4.3.2 Gaussian Partial Information Model

The central component of the Gaussian model is a pool of information particles. Each particle, which can be interpreted as representing the smallest unit of information, is either positive or negative. The positive particles provide evidence in favor of the event $A$, while the negative particles provide evidence against $A$. Therefore, if the overall sum (integral) of the positive particles is larger than that of the negative particles, the event $A$ happens;

otherwise, it does not. Each forecaster, however, observes only the sum of some subset of the particles. Based on this sum, the forecaster makes a probability estimate for $A$. This is made concrete in the following model that represents the pool of information with the unit interval and generates the information particles from a Gaussian process.

> **The Gaussian Model.** Identify the pool of information with the unit interval $S = [0, 1]$. Consider a centered Gaussian process $\{X_B\}$ that is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by the Borel subsets $B \subseteq S$ such that $\mathrm{Cov}\,(X_B, X_{B'}) = |B \cap B'|$. Such a process can be constructed, for example, by considering a standard Brownian motion process $Y(t)$ on $[0, 1]$, and defining $X_B$ as the variation of $Y$ over $B$. Let $A$ denote the event that the sum of all the information is positive: $A := \{X_S > 0\}$. For each $i = 1, \ldots, N$, let $B_i$ be some Borel subset of $S$, and define the corresponding $\sigma$-field as $\mathcal{F}_i := \sigma(X_{B_i})$. Forecaster $i$ then predicts $p_i := \mathbb{E}(\mathbf{1}_A \,|\, \mathcal{F}_i)$.

The Gaussian model can be motivated by recalling the interpreted signal model of Broomell and Budescu (2009). They assume that Forecaster $i$ forms an opinion based on

$$L_i(Z_1, \ldots, Z_r),$$

where each $L_i$ is a linear function of observable quantities or cues $Z_1, \ldots, Z_r$ that determine the outcome of $A$. If the observables (or any linear combination of them) are independent and have small tails, then as $r \to \infty$, the joint distribution of the linear combinations $L_1, \ldots, L_N$ will be asymptotically Gaussian. Therefore, given that the number of cues in a real-world setup is likely to be large, it makes sense to model the forecasters' observations as jointly Gaussian. The remaining component, namely the covariance structure of the joint distribution is then motivated by the partial information framework. Of course, other distributions, such as the $t$-distribution, could be considered. However, given that both the

multivariate and conditional Gaussian distributions have simple forms, the Gaussian model offers potentially the cleanest entry into the issues at hand.

Overall, modeling the forecasters' predictions with a Gaussian distribution is rather common. For instance, Di Bacco et al. (2003) consider a model of two forecasters whose estimated log-odds follow a joint Gaussian distribution. The predictions are assumed to be based on different information sets; hence, the model can be viewed as a partial information model. Unfortunately, as a specialization of the partial information framework, this model is a fairly narrow due to its detailed assumptions and extensive computations. The end result is a rather restricted aggregator of two probability forecasts. On the contrary, the Gaussian model sustains flexibility by specializing the framework only as much as is necessary. The following enumeration provides further interpretation and clarifies which aspects of the model are essential and which have little or no impact.

(i) **Interpretations.** It is not necessary to assume anything about the source of the information. For instance, the information could stem from survey research, records, books, interviews, or personal recollections. All these details have been abstracted away.

(ii) **Information Sets.** The set $B_i$ holds the information used by Forecaster $i$, and the covariance $\mathrm{Cov}\left(X_{B_i}, X_{B_j}\right) = |B_i \cap B_j|$ represents the information overlap between Forecasters $i$ and $j$. Consequently, the complement of $B_i$ holds information not used by Forecaster $i$. No assumption is necessary as to whether this information was unknown to Forecaster $i$ instead of known but not used in the forecast.

(iii) **Pool of Information.** First, the pool of information potentially available to the forecasters is the white noise on $S = [0, 1]$. The role of the unit interval is for the convenient specification of the sets $B_i$. The exact choice is not relevant, and any other set could have been used. The unit interval, however, is a natural starting point that links the information structure to many known results in combinatorics and geometry

(see, e.g., Proposition 4.3.3). Second, there is no sense of time or ranking of information within the pool. Instead, the pool is a collection of information, where each piece of information has an *a priori* equal chance to contribute to the final outcome. Quantitatively, information is parametrized by the length measure on $S$.

(iv) **Invariant Transformations.** From the empirical point of view, the exact identities of the individual sets $B_i$ are irrelevant. All that matters are the covariances $\mathrm{Cov}\left(X_{B_i}, X_{B_j}\right) = |B_i \cap B_j|$. The explicit sets $B_i$ are only useful in the analysis, e.g., when computing the oracular aggregator.

(v) **Scale Invariance.** The model is invariant under rescaling, replacing $S$ by $[0, \lambda]$ and $B_i$ by $\lambda B_i$. Therefore, the actual scale of the model (e.g., the fact that the covariances of the variables $X_B$ are bounded by one) is not relevant.

(vi) **Specific vs. General Model.** A specific model requires a choice of an event $A$ and Borel sets $B_i$. This might be done in several ways: a) by choosing them in advance, according to some criterion; b) estimating the parameters $\mathbb{P}(A)$, $|B_i|$, and $|B_i \cap B_j|$ from data; or c) using a Bayesian model with a prior distribution on the unknown parameters. This paper focuses mostly on a) and b) but discusses c) briefly in Section 5.5. Section 4.4 provides one result, namely Proposition 4.4.2 that holds for any (nonrandom) choices of the sets $B_i$.

(vii) **Choice of Target Event.** There is one substantive assumption in this model, namely the choice of the half-space for the event $A$. Choosing $\{X_S > t\}$ for some $t \in \mathbb{R}$ makes the prior probability equal to $\mathbb{P}(A) = 1 - \Phi(t)$. The current paper defers the analysis of $t \neq 0$ to future work and focuses on the centered model for simplicity. Furthermore, choosing $t = 0$ implies a prior probability $\mathbb{P}(A) = 1/2$, which seems as uninformative as possible and therefore provides a natural starting point. Note that specifying a prior distribution for $A$ cannot be avoided as long as the model depends

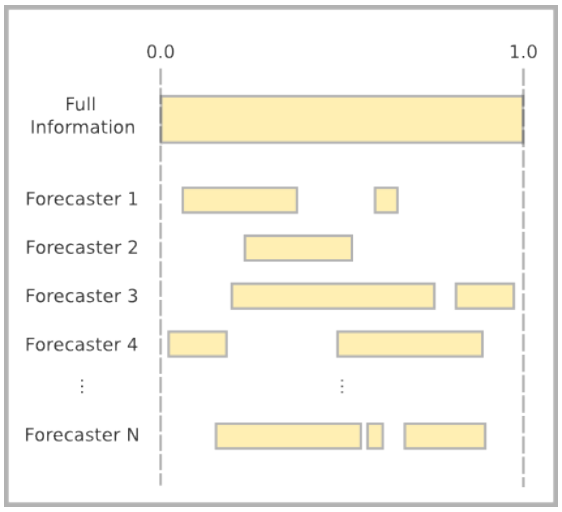



Figure 4.1: Illustration of Information Distribution among $N$ Forecasters. The bars leveled horizontally with Forecaster $i$ represent the information set $B_i$.

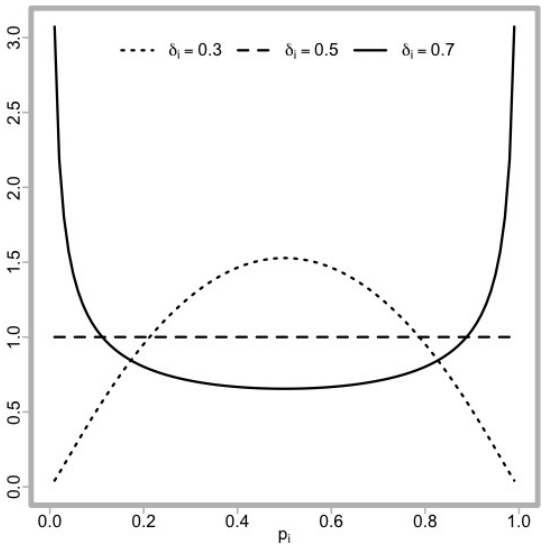



Figure 4.2: Marginal Distribution of $p_i$ under Different Levels of $\delta_i$. The more the forecaster knows, the more the forecasts are concentrated around the extreme points zero and one.

on a probability space. This includes essentially any probability model for forecast aggregation.

### 4.3.3 Preliminary Observations

The Gaussian process exhibits additive behavior that aligns well with the intuition of an information pool. To see this, consider a finite partition of the full information $\{C_{\mathbf{v}} := \cap_{i\in\mathbf{v}} B_i \setminus \cup_{i\notin\mathbf{v}} B_i : \mathbf{v} \subseteq \{1,\ldots,N\}\}$. Each subset $C_{\mathbf{v}}$ represents a set of information particles such that $B_i = \bigcup_{\mathbf{v}\ni i} C_{\mathbf{v}}$ and $X_{B_i} = \sum_{\mathbf{v}\ni i} X_{C_{\mathbf{v}}}$. Therefore $X_B$ can be regarded as the sum of the particles in the subset $B \subseteq S$, and different $X_B$'s relate to each other in a manner that is consistent with this interpretation. The relations among the relevant

variables are summarized by a multivariate Gaussian distribution:

$$\begin{pmatrix} X_S \\ X_{B_1} \\ \vdots \\ X_{B_N} \end{pmatrix} \sim \mathcal{N}\left( \mathbf{0}, \begin{pmatrix} \Sigma_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} = \left( \begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \dots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \dots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \dots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \dots & \delta_N \end{array} \right) \right), \tag{4.2}$$

where $|B_i| = \delta_i$ is the amount of information used by Forecaster $i$, and $|B_i \cap B_j| = \rho_{ij} = \rho_{ji}$ is the amount of information overlap between Forecasters $i$ and $j$. One possible instance of this setup is illustrated in Figure 4.1. Note that $B_i$ does not have to be a contiguous subset of $S$. Instead, each forecaster can use any Borel measurable subset of the full information.

Under the Gaussian model, the sub-matrix $\mathbf{\Sigma}_{22}$ is sufficient for the information structure. Therefore the exact identities of the Borel sets do not matter, and learning about the information among the forecasters is equivalent to estimating a covariance matrix under several restrictions. In particular, if the information in $\mathbf{\Sigma}_{22}$ can be translated into a diagram such as Figure 4.1, the matrix $\mathbf{\Sigma}_{22}$ is called *coherent*. This property is made precise in the following proposition. The proof of this and other propositions are deferred to Appendix A of the Supplementary Material.

**Proposition 4.3.3.** *The overlap structure $\mathbf{\Sigma}_{22}$ is coherent if and only if $\mathbf{\Sigma}_{22} \in$ COR$(N) :=$ conv $\left\{ \boldsymbol{x}\boldsymbol{x}' : \boldsymbol{x} \in \{0,1\}^N \right\}$, where conv$\{\cdot\}$ denotes the convex hull and COR$(N)$ is known as the correlation polytope. It is described by $2^N$ vertices in dimension dim(COR$(N)$) $= \binom{N+1}{2}$.*

The correlation polytope has a very complex description in terms of half-spaces. In fact, complete descriptions of the facets of COR$(N)$ are only known for $N \leq 7$ and conjectured for COR$(8)$ and COR$(9)$ (Ziegler, 2000). Fortunately, previous literature has introduced both linear and semidefinite relaxations of COR$(N)$ (Laurent et al., 1997). Such relaxations

together with modern optimization techniques and sufficient data can be used to estimate the information structure very efficiently. This, however, is not in the scope of this paper and is therefore left for subsequent work.

The multivariate Gaussian distribution (4.2) relates to the forecasts by

$$p_i = \mathbb{P}\left(A|\mathcal{F}_i\right) = \mathbb{P}\left(X_S > 0|X_{B_i}\right) = \Phi\left(\frac{X_{B_i}}{\sqrt{1-\delta_i}}\right). \tag{4.3}$$

The marginal density of $p_i$,

$$m\left(p_i|\delta_i\right) = \sqrt{\frac{1-\delta_i}{\delta_i}} \exp\left\{\Phi^{-1}(p_i)^2\left(1 - \frac{1}{2\delta_i}\right)\right\},$$

has very intuitive behavior: it is uniform on $[0,1]$ if $\delta_i = 1/2$, but becomes unimodal with a minimum (maximum) at $p_i = 1/2$ when $\delta_i > 1/2$ ($\delta_i < 1/2$). As $\delta_i \to 0$, $p_i$ converges to a point mass at $1/2$. On the other hand, as $\delta_i \to 1$, $p_i$ converges to a correct forecast whose distribution has atoms of weight $1/2$ at zero and one. Therefore a forecaster with no information "withdraws" from the problem by predicting a non-informative probability $1/2$ while a forecaster with full information always predicts the correct outcome with absolute certainty. Figure 4.2 illustrates the marginal distribution when $\delta_i$ is equal to $0.3$, $0.5$, and $0.7$.

## 4.4 Probability Extremizing

### 4.4.1 Oracular Aggregator for the Gaussian Model

Recall from Section 4.1.3 that the oracular aggregator is the conditional expectation of $\mathbf{1}_A$ given all the forecasters' information. Under the Gaussian model, this can be emulated with a hypothetical oracle forecaster whose information set is $B' := \bigcup_{i=1}^{N} B_i$. The oracular

aggregator is then nothing more than the probability forecast made by the oracle. That is,

$$p' = \mathbb{P}(A|\mathcal{F}') = \mathbb{P}(X_S > 0|X_{B'}) = \Phi\left(\frac{X_{B'}}{\sqrt{1-\delta'}}\right),$$

where $\delta' = |B'|$. Given that the oracle's information set $B'$ cannot be used to reconstruct the individual sets $\{B_i\}_{i=1}^N$, some potentially relevant information may appear to have been lost. Under the Gaussian model, however, only the total variation over $B'$ is relevant to aggregation. The next proposition shows that $X_{B'}$ contains all the information in $\{X_{B_i}\}_{i=1}^N$ and hence leads to an actual oracular aggregator.

**Proposition 4.4.1.** *The event $A$ is conditionally independent of the collection $\{X_{B_i}\}_{i=1}^N$ given $X_{B'}$*

The oracular aggregator provides a reference point that allows us to identify information structures under which other aggregation techniques perform relatively well. In particular, if an aggregator is likely to be near $p'$ under a given $\mathbf{\Sigma}_{22}$, then that information structure reflects favorable conditions for the aggregator. This ideas is used in the following subsections to develop intuition about probability extremizing.

### 4.4.2 General Information Structure

A probability $p$ is said to be *extremized* by another probability $q$ if and only if $q$ is closer to zero when $p \leq 1/2$ and closer to one when $p \geq 1/2$. This translates to the probit scores as follows: $q$ extremizes $p$ if and only if $\Phi^{-1}(q)$ is on the same side but further away from zero than $\Phi^{-1}(p)$. The amount of (multiplicative) extremization can then be quantified with the *probit extremization ratio* defined as $\alpha(q,p) := \Phi^{-1}(q)/\Phi^{-1}(p)$.

Given that no aggregator can improve upon the oracular aggregator, it provides an ideal reference point for analyzing extremization. This section specifically uses it to study extremizing of $p_{\text{probit}}$ because a) it is arguably more reasonable than the simple average $\bar{p}$;

and b) it is very similar to $p_{\log}$ but results in cleaner analytic expressions. Therefore, of particular interest is the special case $\alpha(p', p_{\text{probit}}) = P' \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} P_i\right)$, where $P' = \Phi^{-1}(p')$. From now on, unless otherwise stated, this expression is referred simply with $\alpha$. Therefore, the probit opinion pool $p_{\text{probit}}$ requires extremization if and only if $\alpha > 1$, and the larger $\alpha$ is, the more $p_{\text{probit}}$ should be extremized.

Note that $\alpha$ is a random quantity that spans the entire real line; that is, it is possible to find a set of forecasts and an information structure for any possible value of $\alpha \in \mathbb{R}$. Evidently, extremizing is not guaranteed to always improve $p_{\text{probit}}$. To understand when extremizing is likely to be beneficial, the following proposition provides the probability distribution of $\alpha$.

**Proposition 4.4.2.** *The law of the extremization ratio $\alpha$ is a Cauchy with parameters $x_0$ and $\gamma$, where the location parameter $x_0$ is at least one, equality occurring only when $\delta_i = \delta_j$ for all $i \neq j$. Consequently, if $\delta_i \neq \delta_j$ for some $i \neq j$, then the probability that $p_{\text{probit}}$ requires extremizing $\mathbb{P}\left(\alpha > 1 | \boldsymbol{\Sigma}_{22}, \delta'\right)$ is strictly greater than $1/2$.*

This proposition shows that, on any non-trivial problem, a small perturbation in the direction of extremizing is more likely to improve $p_{\text{probit}}$ than to degrade it. This partially explains why extremizing aggregators perform well on large sets of real-world prediction problems. It may be unsurprising after the fact, but the forecasting literature is still full of articles that perform probability averaging without extremizing. The next two subsections examine special cases in which more detailed computations can be performed.

### 4.4.3 Zero and Complete Information Overlap

If the forecasters use the same information, i.e., $B_i = B_j$ for all $i \neq j$, their forecasts are identical, $p' = p'' = p_{\text{probit}}$, and no extremization is needed. Therefore, given that the oracular aggregator varies smoothly over the space of information structures, averaging techniques, such as $p_{\text{probit}}$, can be expected to work well when the forecasts are based

on very similar sources of information. This result is supported by the fact that the measurement error framework, which essentially describes the forecasters as making numerous small mistakes while applying the same procedure to the same data (see Section 4.2.2), results in averaging-based aggregators.

If, on the other hand, the forecasters have zero information overlap, i.e., $|B_i \cap B_j| = 0$ for all $i \neq j$, the information structure $\mathbf{\Sigma}_{22}$ is diagonal and

$$p' = p'' = \Phi\left(\frac{\sum_{i=1}^{N} X_{B_i}}{\sqrt{1 - \sum_{i=1}^{N} \delta_i}}\right),$$

where the identities $\delta' = \sum_{i=1}^{N} \delta_i$ and $X_{B'} = \sum_{i=1}^{N} X_{B_i}$ result from the additive nature of the Gaussian process (see Section 4.3.3). This aggregator can be described in two steps: First, the numerator conducts voting, or range voting to be more specific, where the votes are weighted according to the importance of the forecasters' private information. Second, the denominator extremizes the consensus according to the total amount of information in the group. This clearly leads to very extreme forecasts. Therefore more extreme techniques can be expected to work well when the forecasters use widely different information sets.

The analysis suggests a spectrum of aggregators indexed by the information overlap: the optimal aggregator undergoes a smooth transformation from averaging (low extremization) to voting (high extremization) as the information overlap decreases from complete to zero overlap. This observation gives qualitative guidance in real-world settings where the general level of overlap can be said to be high or low. For instance, predictions from forecasters working in close collaboration can be averaged while predictions from forecasters strategically accessing and studying disjoint sources of information should be aggregated via more extreme techniques such as voting. See Parunak et al. 2013 for a discussion of voting-like techniques. For a concrete illustration, recall Example 4.3.1 where the optimal aggregate changes from $2/3$ (high information overlap) to $4/5$ (low information overlap).

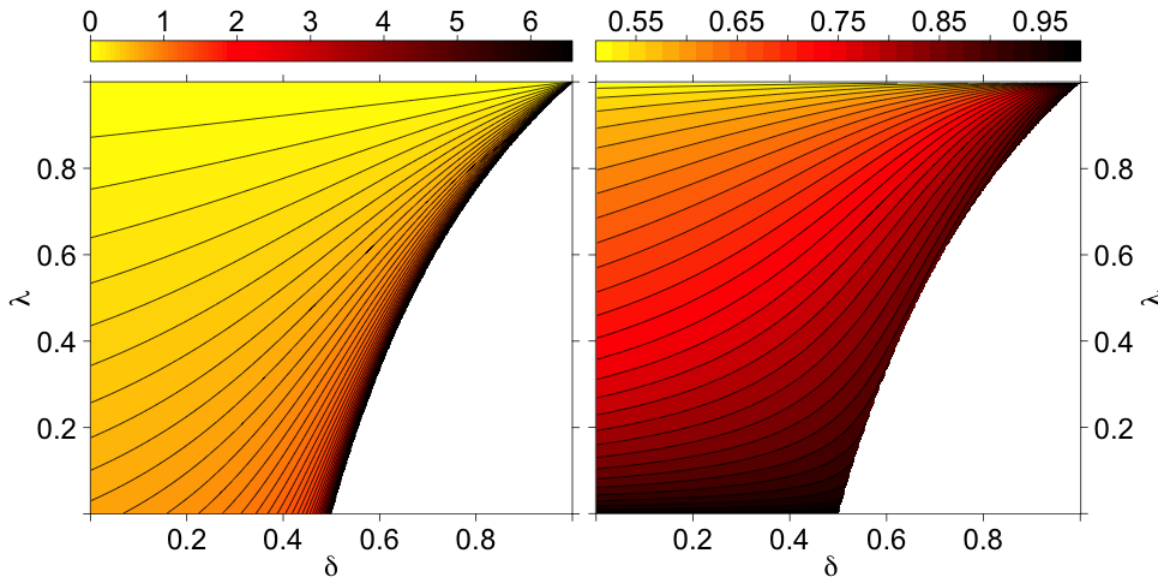
### 4.4.4 Partial Information Overlap

To analyze the intermediate scenarios with partial information overlap among the forecasters, it is helpful to reduce the number of parameters in $\mathbf{\Sigma}_{22}$. A natural approach is to assume compound symmetry, where the information sets have the same size and that the amount of pairwise overlap is constant. More specifically, let $|B_i| = \delta$ and $|B_i \cap B_j| = \lambda\delta$, where $\delta$ is the amount of information used by each forecaster and $\lambda$ is the overlapping proportion of this information. The resulting information structure is $\mathbf{\Sigma}_{22} = \mathbf{I}_N(\delta - \lambda\delta) + \mathbf{J}_N\lambda\delta$, where $\mathbf{I}_N$ is the identity matrix and $\mathbf{J}_N$ is $N \times N$ matrix of ones. It is coherent if and only if

$$\delta \in [0, 1] \qquad \text{and} \qquad \lambda|\delta \in \left[\max\left\{\frac{N - \delta^{-1}}{N - 1}, 0\right\}, 1\right]. \tag{4.4}$$

See Appendix A of the Supplementary Material for the derivation of these constraints.

Under these assumptions, the location parameter of the Cauchy distribution of $\alpha$ simplifies to $x_0 = N/(1+(N-1)\lambda)\sqrt{(1-\delta)/(1-\delta')}$. Of particular interest is to understand how this changes as a function of the model parameters. The analysis is somewhat hindered by the unknown details of the dependence between $\delta'$ and the other parameters $N$, $\delta$, and $\lambda$. However, given that $\delta'$ is defined as $\delta' = |\cup_{i=1}^{N} B_i|$, its value increases in $N$ and $\delta$ but decreases in $\lambda$. In particular, as $\delta \to 1$, the value of $\delta'$ converges to one at least as fast as $\delta$ because $\delta' \geq \delta$. Therefore the term $\sqrt{(1-\delta)/(1-\delta')}$ and, consequently, $x_0$ increase in $\delta$. Similarly, $x_0$ can be shown to increase in $N$ but to decrease in $\lambda$. Therefore $x_0$ and $\delta'$ move together, and the amount of extremizing can be expected to increase in $\delta'$. As the Cauchy distribution is symmetric around $x_0$, the probability $\mathbb{P}(\alpha > 1|\mathbf{\Sigma}_{22})$ behaves similarly to $x_0$ and also increases in $\delta'$. Figure 4.3 illustrates these relations by plotting both $\log(x_0)$ and $\mathbb{P}(\alpha > 1|\mathbf{\Sigma}_{22})$ for $N = 2$ forecasters under all plausible combinations of $\delta$ and $\lambda$. The white space collects all pairs $(\delta, \lambda)$ that do not satisfy (4.4) and hence represent incoherent information structures. Note that the results are completely general for the two-forecaster




(a) $\log(x_0)$ (b) $\mathbb{P}(\alpha > 1|\mathbf{\Sigma}_{22})$

Figure 4.3: Extremization Ratio under Symmetric Information. The amount of extremizing $\alpha$ follows a Cauchy$(x_0, \gamma)$, where $x_0$ is a location parameter and $\gamma$ is a scale parameter. This figure considers $N = 2$ because in this case $\delta'$ is uniquely determined by $\mathbf{\Sigma}_{22}$.

case, apart from the assumption $\delta_1 = \delta_2$. Relaxing this assumption does not change the qualitative nature of the results.

The total amount of information used by the forecasters $\delta'$, however, does not provide a full explanation of extremizing. Information diversity is an important yet separate determinant. To see this, observe that fixing $\delta'$ to some constant defines a curve $\lambda = 2 - \delta'/\delta$ on the two plots in Figure 4.3. For instance, letting $\delta' = 1$ gives the boundary curve on the right side of each plot. This curve then shifts inwards and rotates slightly counterclockwise as $\delta'$ decreases. At the top end of each curve all forecasters use the total information, i.e., $\delta = \delta'$ and $\lambda = 1.0$. At the bottom end, on the other hand, the forecasters partition the total information and have zero overlap, i.e., $\delta = \delta'/2$ and $\lambda = 0.0$. Given that moving down along these curves simultaneously increases information diversity and $x_0$, both information diversity and the total amount of information used by the forecasters are important yet sep-

arate determinants of extremizing. This observation can guide practitioners towards proper extremization because many application specific aspects are linked to these two determinants. For instance, extremization can be expected to increase in the number of forecasters, subject-matter expertise, and human diversity, but to decrease in collaboration, sharing of resources, and problem difficulty.

## 4.5 Probability Aggregation

### 4.5.1 Revealed Aggregator for the Gaussian Model

Recall the multivariate Gaussian distribution (4.2) and collect all $X_{B_i} = \Phi^{-1}(p_i)\sqrt{1-\delta_i}$ into a column vector $\boldsymbol{X} = (X_{B_1}, X_{B_2}, \ldots, X_{B_N})'$. If $\boldsymbol{\Sigma}_{22}$ is a coherent overlap structure and $\boldsymbol{\Sigma}_{22}^{-1}$ exists, then the revealed aggregator under the Gaussian model is

$$p'' = \mathbb{P}\left(A|\mathcal{F}''\right) = \mathbb{P}\left(X_S > 0|\boldsymbol{X}\right) = \Phi\left(\frac{\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}}{\sqrt{1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}}}\right). \tag{4.5}$$

Applying (4.5) in practice requires an estimate of $\boldsymbol{\Sigma}_{22}$. If the forecasters make predictions about multiple events, it may be possible to model the different prediction tasks with a hierarchical structure and estimate a fully general form of $\boldsymbol{\Sigma}_{22}$. This can be formulated as a constrained (semi-definite) optimization problem, which, as was mentioned in Section 4.3.3, is left for subsequent work. Such estimation, however, requires the results of a large multi-prediction experiment which may not always be possible in practice. Often only a single prediction per forecaster is available. Consequently, accurate estimation of the fully general information structure becomes difficult. This motivates the development of aggregation techniques for a single event. Under the Gaussian model, a standard approach is to assume a covariance structure that involves fewer parameters. The next subsection discusses a natural and non-informative choice.

### 4.5.2 Symmetric Information

This subsection assumes a type of exchangeability among the forecasters. While this is somewhat idealized, it is a reasonable choice in a low-information environment where there is no historical or self-report data to distinguish the forecasters. The averaging aggregators described in Section 4.2, for instance, are symmetric. Therefore, to the extent that they reflect an underlying model, the model assumes exchangeability. Under the Gaussian model, exchangeability suggests the compound symmetric information structure discussed in Section 4.4.4. This structure holds if, for example, the forecasters use information sources sampled from a common distribution. The resulting revealed aggregator takes the form

$$p''_{cs} = \Phi\left(\frac{\frac{1}{(N-1)\lambda+1}\sum_{i=1}^{N} X_{B_i}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}}\right), \tag{4.6}$$

where $X_{B_i} = \Phi^{-1}(p_i)\sqrt{1-\delta}$ for all $i = 1, \ldots, N$.

Given these interpretations, it may at first seem surprising that the values of $\delta$ and $\lambda$ can be estimated in practice. Intuitively, the estimation relies on two key aspects of the model: a) a better-informed forecast is likely to be further away from the non-informative prior (see Figure 4.2); and b) two forecasters with high information overlap are likely to report very similar predictions. This provides enough leverage to estimate the information structure via the maximum likelihood method. Complete details for this are provided in Appendix B of the Supplementary Material. Besides exchangeability, $p''_{cs}$ is based on very different modeling assumptions than the averaging aggregators. The following proposition summarizes some of its key properties.

**Proposition 4.5.1.** $(i)$ *The probit extremization ratio between* $p''_{cs}$ *and* $p_{\text{probit}}$ *is given by the non-random quantity* $\alpha(p''_{cs}, p_{\text{probit}}) = \gamma\sqrt{1-\delta}/\sqrt{1-\delta\gamma}$*, where* $\gamma = N/((N-1)\lambda+1)$*,*

$(ii)$ *$p''_{cs}$ extremizes $p_{\text{probit}}$ as long as $p_i \neq p_j$ for some $i \neq j$, and*

$(iii)$ *$p''_{cs}$ can leave the convex hull of the individual probability forecasts.*

Proposition 4.5.1 suggests that $p''_{cs}$ is appropriate for combining probability forecasts of a single event. This is illustrated on real-world forecasts in the next subsection. The goal is not to perform a thorough data analysis or model evaluation, but to demonstrate $p''_{cs}$ on a simple example.

### 4.5.3 Real-World Forecasting Data

Probability aggregation appears in many facets of real-world applications, including weather forecasting, medical diagnosis, estimation of credit default, and sports betting. This section, however, focuses on predicting global events that are of particular interest to the Intelligence Advanced Research Projects Activity (IARPA). Since 2011, IARPA has posed about 100-150 question per year as a part of its ACE forecasting tournament. Among the participating teams, the Good Judgment Project (GJP) (Ungar et al. 2012; Mellers et al. 2014) has emerged as the clear winner. The GJP has recruited thousands of forecasters to estimate probabilities of the events specified by IARPA. The forecasters are told that their predictions are assessed using the Brier score (see Section 4.1.2). In addition to receiving \$150 for meeting minimum participation requirements that do not depend on prediction accuracy, the forecasters receive status rewards for good performance via leader-boards displaying Brier scores for the top 20 forecasters. Every year the top 1% of the forecasters are selected to the elite group of "super-forecasters". Note that, depending on the details of the reward structure, such a competition for rank may eliminate the truth-revelation property of proper scoring rules (see, e.g., Lichtendahl Jr and Winkler 2007).

This subsection focuses on the super-forecasters in the second year of the tournament. Given that these forecasters were elected to the group of super-forecasters based on the first year, their forecasts are likely, but not guaranteed, to be relatively good. The group

involves 44 super-forecasters collectively making predictions about 123 events, of which 23 occurred. For instance, some of the questions were: "Will France withdraw at least 500 troops from Mali before 10 April 2013?", and "Will a banking union be approved in the EU council before 1 March 2013?". Not every super-forecaster made predictions about every event. In fact, the number of forecasts per event ranged from 17 to 34 forecasts, with a mean of 24.2 forecasts. To avoid infinite log-odds and probit scores, extreme forecasts $p_i = 0$ and 1 were censored to $p_i = 0.001$ and $0.999$, respectively.

In this section aggregation is performed one event at a time without assuming any other information besides the probability forecasts themselves. This way any performance improvements reflect better fit of the underlying model and the aggregator's relative advantage in forecasting a single event. Aggregation accuracy is measured with the mean Brier score (BS): Consider $K$ events and collect all $N_k$ probability forecasts for event $A_k$ into a vector $\boldsymbol{p}_k \in [0,1]^{N_k}$. Then, BS for aggregator $g : [0,1]^{N_k} \to [0,1]$ is

$$\text{BS} = \frac{1}{K} \sum_{k=1}^{K} \left( g(\boldsymbol{p}_k) - \mathbf{1}_{A_k} \right)^2 .$$

This score is defined on the unit interval with lower values indicating higher accuracy. For a more detailed performance analysis, it decomposes into three additive components: reliability (REL), resolution (RES), and uncertainty (UNC). This assumes that the aggregate forecast $g(\boldsymbol{p}_k)$ for all $k$ can only take discrete values $f_j \in [0,1]$ with $j = 1, \ldots, J$. Let $n_j$ be the number of times $f_j$ occurs, and denote the empirical frequency of the corresponding events with $o_j$. Let $\bar{o}$ be the overall empirical frequency of occurrence, i.e., $\bar{o} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{A_k}$. Then,

$$\begin{aligned} \text{BS} &= \text{REL} - \text{RES} + \text{UNC} \\ &= \frac{1}{K} \sum_{j=1}^{J} n_j (f_j - o_j)^2 - \frac{1}{K} \sum_{j=1}^{J} n_j (o_j - \bar{o})^2 + \bar{o}(1 - \bar{o}). \end{aligned}$$

Table 4.1: The Mean Brier Scores (BS) with Its Three Components, Reliability (REL), Resolution (RES), and Uncertainty (UNC), for Different Aggregators.

| Aggregator | BS | REL | RES | UNC |
|---|---|---|---|---|
| $\bar{p}$ | 0.132 | 0.026 | 0.045 | 0.152 |
| $p_{\log}$ | 0.128 | 0.025 | 0.048 | 0.152 |
| $p_{\text{probit}}$ | 0.128 | 0.023 | 0.047 | 0.152 |
| $p''_{cs}$ | 0.123 | 0.020 | 0.049 | 0.152 |

In this decomposition low REL represents good calibration. If a calibrated aggregate is also confident, it exhibits high RES. Therefore the combination of good calibration and high confidence leads to low BS. The corresponding forecasts are likely to be very close to $0$ and $1$, which is more useful to the decision-maker than the naive forecast $\bar{o}$. The final term UNC equals the BS for $\bar{o}$ and hence provides a reference point for interpreting the performance of the aggregator.

Table 4.1 presents results for $\bar{p}$, $p_{\log}$, $p_{\text{probit}}$, and $p''_{cs}$ under the super-forecaster data. Empirical approaches were not considered for two reasons: a) they do not reflect an actual model of forecasts; and b) they require a training set with known outcomes and hence cannot be applied to a single event. Overall, $\bar{p}$ presents the worst performance. Given that $p_{\text{probit}}$ and $p_{\log}$ are very similar, it is not surprising that they have almost identical scores. The revealed aggregator $p''_{cs}$ is both the most resolved and calibrated, thus achieving the lowest BS among all the aggregators. This is certainly an encouraging result. It is important to note that $p''_{cs}$ is only the first attempt at partial information aggregation. More elaborate information structures and estimation procedures, such as shrinkage estimators, are very likely to lead to many further improvements.

## 4.6 Summary and Discussion

This paper introduced a probability model for predictions made by a group of forecasters. The model allows for interpretation of some of the existing work on forecast aggregation and also clarifies empirical approaches such as the *ad hoc* practice of extremization. The general model is more plausible on the micro-level than any other model has been to date. Under this model, some general results were provided. For instance, the *oracular* aggregate, which uses all the forecasters' information (Proposition 4.4.1), is more likely to be more extreme than one of the common benchmark aggregates, namely $p_{\text{probit}}$ (Proposition 4.4.2). Even though no real world aggregator has access to all the information of the oracle, this result explains why extremization is almost certainly called for. More detailed analyses were performed under several specific model specifications such as zero and complete information overlap (Section 4.4.3), and fully symmetric information (Section 4.4.4). Even though the zero and complete information overlap models are not realistic, except under a very narrow set of circumstances, they form logical extremes that illustrate the main drivers of good aggregation. The symmetric model is somewhat more realistic. It depends only on two parameters and therefore allows us to visualize the effect of model parameters on the optimal amount of extremization (Figure 4.3). Finally, the *revealed* aggregator, which is the best in-practice aggregation under the partial information model, was discussed. The discussion provided a general formula for this aggregator (Equation 4.5) as well as its specific formula under symmetric information (Equation 4.6). The specific form was applied to real-world forecasts of one-time events and shown to outperform other model-based aggregators.

It is interesting to relate our discussion to the many empirical studies conducted by the Good Judgment Project (GJP) (see Section 7.5). Generally extremizing has been found to improve the average aggregates (Mellers et al., 2014; Satopää et al., 2014,?). The average forecast of a team of super-forecasters, however, often requires very little or no extremizing.

This can be explained as follows. The super-forecasters are highly knowledgeable (high $\delta$) individuals who work in groups (high $\rho$ and $\lambda$). Therefore, in Figure 4.3 they are situated around the upper-right corners where almost no extremizing is required. In other words, there is very little left-over information that is not already used in each forecast. Their forecasts are highly convergent and are likely to be already very near the oracular forecast. The GJP forecast data also includes self-assessments of expertise. Not surprisingly, the greater the self-assessed expertise, the less extremizing appears to have been required. This is consistent with our interpretation that high values of $\delta$ and $\lambda$ suggest lower extremization.

The partial information framework offers many directions for future research. One involves estimation of parameters. In principle, $|B_i|$ can be estimated from the distribution of a reasonably long probability stream. Similarly, $|B_i \cap B_j|$ can be estimated from the correlation of the two parallel streams. Estimation of higher order intersections, however, seems more dubious. In some cases the higher order intersections have been found to be irrelevant to the aggregation procedure. For instance, DeGroot and Mortera (1991) show that it is enough to consider only the pairwise conditional (on the truth) distributions of the forecasts when computing the optimal weights for a linear opinion pool. Theoretical results on the significance or insignificance of higher order intersections under the partial information framework would be desirable.

Another promising avenue is the Bayesian approach. In many applications with small or moderately sized datasets, Bayesian methods have been found to be superior to the likelihood-based alternatives. Therefore, given that the number of forecasts on a single event is typically quite small, a Bayesian approach is likely to improve the predictions of one-time events. Currently, we have work in progress analyzing a Bayesian model but there are many, many reasonable priors on the information structures. This avenue should certainly be pursued further, and the results tested against other high performing aggregators.

## 4.7 Acknowledgments

This research was supported in part by NSF grant # DMS-1209117 and a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. The authors would like to thank Edward George and Shane Jensen for helpful discussions.

# 5

# Partial Information Framework: Model-Based Aggregation of Estimates from Diverse Information Sources*


## Abstract

Prediction polling is an increasingly popular form of crowdsourcing in which multiple participants estimate the probability or magnitude of some future event. These estimates are then aggregated into a single forecast. Historically, randomness in scientific estimation has been generally assumed to arise from unmeasured factors which are viewed as measurement noise. However, when combining subjective estimates, heterogeneity stemming from differences in the participants' information is often more important than measurement noise. This paper formalizes information diversity as an alternative source of such heterogeneity and introduces a novel modeling framework that is particularly well-suited for prediction polls. A practical specification of this framework is proposed and applied to the task of aggregating probability and point estimates from two real-world prediction polls. In both cases our model outperforms standard measurement-error-based aggregators,


*Joint work with Shane T. Jensen, Robin Pemantle, and Lyle H. Ungar

hence providing evidence in favor of information diversity being the more important source of heterogeneity.

## 5.1 Introduction

Past literature has distinguished two types of polling: prediction and opinion polling. In broad terms, an opinion poll is a survey of public opinion, whereas a prediction poll involves multiple agents collectively predicting the value of some quantity of interest (Goel et al., 2010; Mellers et al., 2014). For instance, consider a presidential election poll. An opinion poll typically asks the voters who they will vote for. A prediction poll, on the other hand, could ask which candidate they think will win in their state. A liberal voter in a dominantly conservative state is likely to answer differently to these two questions. Even though opinion polls have been the dominant focus historically, prediction polls have become increasingly popular in the recent years, due to modern social and computer networks that permit the collection of a large number of responses both from human and machine agents. This has given rise to crowdsourcing platforms, such as MTurk and Witkey, and many companies, such as Myriada, Lumenogic, and Inkling, that have managed to successfully capitalize on the benefits of collective wisdom.

This paper introduces statistical methodology designed specifically for the rapidly growing practice of prediction polling. The methods are illustrated on real-world data involving two common types of responses, namely probability and point forecasts. The probability forecasts were collected by the Good Judgment Project (GJP) (Ungar et al. 2012; Mellers et al. 2014) as a means to estimate the likelihoods of international political future events deemed important by the Intelligence Advanced Research Projects Activity (IARPA). Since its initiation in 2011, the project has recruited thousands of forecasters to make probability estimates and update them whenever they felt the likelihoods had changed. To illustrate, Figure 5.1 shows the forecasts for one of these events. This example involves $522$ fore-

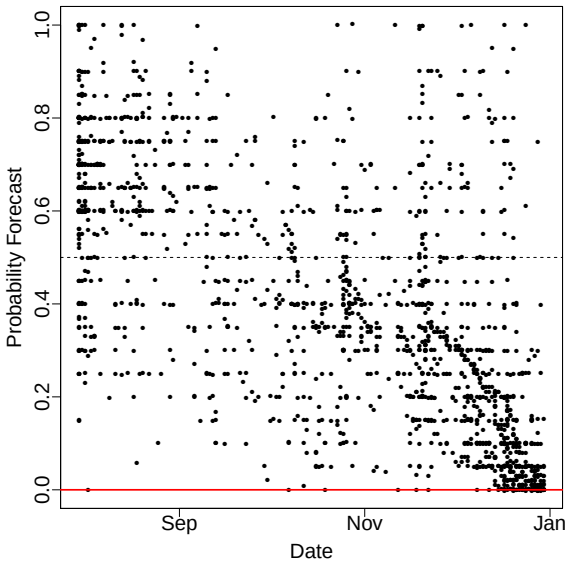



Figure 5.1: Probability forecasts of the event "Will Moody's issue a new downgrade on the long-term ratings for any of the eight major French banks between 30 July 2012 and 31 December 2012?" The points have been jittered slightly to make overlaps visible.

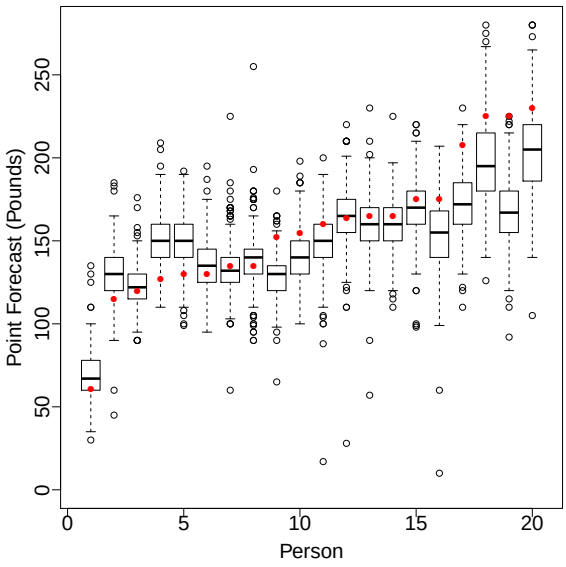



Figure 5.2: Point forecasts of the weights of 20 different people. The boxplots have been sorted to increase in the true weights (red dots). Some extreme values were omitted for the sake of clarity.

casters making a total of $1,669$ predictions between 30 July 2012 and 30 December 2012 when the event finally resolved as "No" (represented by the red line at $0.0$). In general, the forecasters reported updates very infrequently. Furthermore, not all forecasters made probability estimates for all the events, making the dataset very sparse. The point forecasts for our second application were collected by Moore and Klein (2008) who recruited $416$ undergraduates from Carnegie Mellon University to guess the weights of $20$ people based on a series of pictures. This is an experimental setup where each participant was required to respond to all the questions, leading to a fully completed dataset. The responses are illustrated in Figure 5.2 that shows the boxplots of the forecasters' guesses for each of the

20 people. The red dots represent the corresponding true weights.

Once the predictions have been collected, they are typically combined into a single consensus forecast for the sake of decision-making and improved accuracy. Unfortunately, this can be done in many different ways, and the final combination rule can largely determine the out-of-sample performance. The past literature distinguishes two broad approaches to forecast aggregation: empirical aggregation and model-based aggregation. Empirical aggregation is by far the more widely studied approach; see, e.g., stacking (Breiman, 1996), Bayes model averaging (Raftery et al., 1997), linear opinion pools (DeGroot and Mortera, 1991), and extremizing aggregators (Ranjan and Gneiting, 2010; Satopää et al., 2014). All these methods are akin to machine learning in a sense that they first learn the aggregator based on a training set of past forecasts of known outcomes and then use that aggregator to combine future forecasts of unknown outcomes. Unfortunately, in a prediction polling setup, constructing such a training set requires a lot of effort and time on behalf of the forecasters and the polling agent. Therefore a training set is often not available. Instead, the participants are typically handed a single questionnaire that simultaneously inquires about their predictions of one or more unknown outcomes. This leads to a dataset consisting only of forecasts, which means that empirical aggregation cannot be applied.

Fortunately, model-based aggregation can be performed even when prior knowledge of outcomes is not available. This approach begins by proposing a plausible probability model for the source of heterogeneity among the forecasts, that is, for how and why the forecasts differ from the target outcome. Under this assumed forecast-outcome link, it is then possible to construct an optimal aggregator that can be applied directly to the forecasts without learning the aggregator first from a separate training set. Given this broad applicability, the current paper focuses only on the model-based approach. In particular, outcomes are not assumed available for aggregation at any point in the paper. Instead, aggregation is performed solely based on forecasts, leaving all empirical techniques well outside the

scope of the paper.

Historically, potentially due to early forms of data collection, model-based aggregation has considered measurement error as the main source of forecast heterogeneity. This choice motivates aggregators with central tendency such as the (weighted) average, median, and so on. Intuitively, measurement error may be reasonable in modeling repeated estimates from a single instrument. However, it is unlikely to hold in prediction polling, where the estimates arise from multiple, often widely different sources. It is also known that a non-trivial weighted average is not the optimal aggregator (in terms of the expected quadratic and many other loss functions) under any joint distribution of the outcome and its (conditionally unbiased) forecasts (Dawid et al., 1995; Ranjan and Gneiting, 2010; Satopää and Ungar, 2015). This questions the role of measurement error in model-based aggregation and highlights the need for a different source of forecast heterogeneity.

The main contribution of this paper is a new source of forecast heterogeneity, called *information diversity*, that explains variation by differences in the information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles about a company may use separate parts of the information and hence report differing predictions on the company's future revenue. Such diversity forms the basis of a novel modeling framework known as the *partial information framework*. Theory behind this framework was originally introduced for probability forecasts by Satopää et al. (2015); though their specification is somewhat restrictive for empirical applications. The current paper generalizes the framework beyond probability forecast and removes all unnecessary assumptions, leading to a new specification that is more appropriate for practical applications. This specification allows the decision-maker to build models for different types of forecast-outcome pairs, such as probability forecasts of binary events or point forecasts of real-valued outcomes. Each such model motivates and describes an explicit joint distribution for the target outcome and its forecasts. The optimal aggregator under

this joint distribution is available and serves as a more principled model-based alternative to the usual (weighted) average or median.

The paper is structured as follows. Section 5.2 first describes the partial information framework at its most general level and then introduces a practical specification of the framework. The section ends with a brief review of previous work on model-based aggregation. Section 5.3 derives a general procedure that guides efficient estimation of the information structure among the forecasters. Section 5.4 illustrates on real-world data how specific models within the framework can be constructed and applied. In particular, the models are derived and evaluated on probability and point forecasts from the two prediction polls discussed above. Overall, the resulting partial information aggregators achieve a noticeable performance improvement over the common measurement-error-based aggregators, suggesting that information diversity is the more appropriate model of forecast heterogeneity. Finally, Section 5.5 concludes with a summary and discussion of future research.

## 5.2 Model-Based Aggregation

### 5.2.1 Bias and Noise

Consider $N$ forecasters and suppose forecaster $j$ predicts $X_j$ for some quantity of interest $Y$. For instance, in our weight estimation example $Y$ is the true weight of a person and $X_j$ is the guess given by the $j$th undergraduate. In our probability forecasting application, on the other hand, $Y$ is binary, reflecting whether the event happens or not, and $X_j \in [0, 1]$ is a probability forecast for its occurrence. This section, however, avoids such application specific choices and treats $Y$ and $X_j$ as generic random variables. In general, prediction $X_j$ is nothing but an estimator of $Y$. Therefore, as is the case with all estimators, its deviation from the truth can be broken down into two components: bias and noise. On the theoretical level, these two components can be separated and hence are often addressed by different

mechanisms. This suggests a two-step approach to forecast aggregation: i) eliminate any bias in the forecasts, and ii) combine the unbiased forecasts.

Historically, bias in human judgment has been extensively studied in the psychology literature (for reviews, see Lichtenstein et al. 1977; Yates 1990; Keren 1991). This bias often exhibits well-known patterns (see, e.g., the easy-hard effect in Lichtenstein and Fischhoff 1977; Juslin 1993), and many authors have proposed both cognitive and motivational models to explain it (Koriat et al., 1980; Kruglanski, 1990; Soll, 1996; Moore and Healy, 2008). These models and other results in this popular area of research suggest ways for ex-ante bias reduction. Such techniques, however, are not in the scope of this paper. Instead, the focus here is on noise reduction and hence specifically on developing methodology for the second step in the overall process of forecast aggregation. In particular, Section 5.2.2 describes our new framework for modeling the noise component. This is then compared in Section 5.2.3 to previous noise models. These models make different assumptions about the way the unbiased forecasts relate to the target outcome and hence motivate very different classes of model-based aggregators.

## 5.2.2 Partial Information Framework

### 5.2.2.1 General Framework

The partial information framework assumes that $Y$ and $X_j$ are measurable under some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability measure $\mathbb{P}$ provides a non-informative yet proper prior on $Y$ and reflects the *basic information* known to all forecasters. Such a prior has been discussed extensively in the economics and game theory literature where it is usually known as the *common prior*. Even though this is a substantive assumption in the framework, specifying a prior distribution cannot be avoided as long as the model depends on a probability space. This includes essentially any probability model for forecast aggregation. How the prior is incorporated depends on the problem context: it can be chosen

explicitly by the decision-maker, computed based on past observations of $Y$, or estimated directly from the forecasts.

The principal $\sigma$-field $\mathcal{F}$ can be interpreted as all the possible information that can be known about $Y$. On top of the basic information reflected in the prior, the $j$th forecaster uses some personal partial information set $\mathcal{F}_j \subseteq \mathcal{F}$ and predicts $X_j = \mathbb{E}(Y \mid \mathcal{F}_j)$. Therefore $\mathcal{F}_i \neq \mathcal{F}_j$ if $X_i \neq X_j$, and forecast heterogeneity stems purely from *information diversity*. Note, however, that if forecaster $j$ uses a simple rule, $\mathcal{F}_j$ may not be the full $\sigma$-field of information available to the forecaster but rather a smaller $\sigma$-field corresponding to the information used by the rule. Furthermore, if two forecasters have access to the same $\sigma$-field, they may decide to use different sub-$\sigma$-fields, leading to different predictions. This is particularly salient in our weight estimation example where each forecaster has access to the exact same information, namely the picture of the person, but can choose to use different subsets of this information. Therefore, information diversity does not only arise from differences in the available information, but also from how the forecasters decide to use it. This general point of view was motivated in Satopää et al. (2015) with simple examples that illustrate how the optimal aggregate is not well-defined without assumptions on the information structure among the forecasters.

Satopää et al. (2015) also show that $X_j = \mathbb{E}(Y \mid \mathcal{F}_j)$ is precisely the same as having a calibrated (sometimes also known as reliable) forecast, that is, $X_j = \mathbb{E}(Y|X_j)$. Therefore the form $X_j = \mathbb{E}(Y|\mathcal{F}_j)$ arises directly from the existence of an underlying probability model and calibration. Overall, calibration $X_j = \mathbb{E}(Y|X_j)$ has been widely discussed in the statistical and meteorological forecasting literature (see, e.g., Dawid et al. 1995; Ranjan and Gneiting 2010; Jolliffe and Stephenson 2012, Section 7.2.2.), with traces at least as far back as Murphy and Winkler (1987b). Given that the condition $X_j = \mathbb{E}(Y|X_j)$ depends on the probability measure $\mathbb{P}$, it should be referred to as $\mathbb{P}$-calibration when the choice of the probability measure needs to be emphasized. This dependency shows the

main conceptual difference between $\mathbb{P}$-calibration and the notion of empirical calibration (Dawid 1982; Foster and Vohra 1998; and many others). However, as was pointed out by Dawid et al. (1995), these two notions can be expressed in formally identical terms by letting $\mathbb{P}$ represent the limiting joint distribution of the forecast-outcome pairs.

In practice researchers have discovered many calibrated subpopulations of experts, such as meteorologists (Murphy and Winkler, 1977a,b), experienced tournament bridge players (Keren, 1987), and bookmakers (Dowie, 1976). Generally, calibration can be improved through team collaboration, training, tracking (Mellers et al., 2014), performance feedback (Murphy and Daan, 1984), representative sampling of target events (Gigerenzer et al., 1991; Juslin, 1993), or by evaluating the forecasters' performance under a loss function that is minimized by the conditional expectation of $Y$, given the forecaster's information (Banerjee et al., 2005). If one is nonetheless left with uncalibrated forecasts, they can be calibrated ex-ante as follows. First, consider some (possibly uncalibrated) forecasts $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_N)'$ defined on $(\Omega, \mathcal{F})$. Choose some distribution $\mathbb{Q}$ for $(Y, \tilde{\mathbf{X}})$. For instance, Dawid et al. 1995 suggest first choosing a distribution $\mathbb{Q}$ for $\tilde{\mathbf{X}}$ and then setting $\mathbb{Q}(Y, \tilde{\mathbf{X}}) = \Psi(\tilde{\mathbf{X}})\mathbb{Q}(\tilde{\mathbf{X}})$, where $\Psi$ is an arbitrary aggregator (such as the average of probability forecasts of a binary event) acting as $\mathbb{Q}(Y|\tilde{\mathbf{X}})$. Alternatively, one may search for an appropriate $\mathbb{Q}$ in the large literature of quantitative psychology. Regardless how $\mathbb{Q}$ is constructed, however, the calibrated version of $\tilde{X}_j$ is $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j)$. This forecast is $\mathbb{Q}$-calibrated and can be written as $\mathbb{E}_{\mathbb{Q}}(Y|\mathcal{F}_j)$, where $\mathcal{F}_j = \sigma(\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j))$ is the $\sigma$-field generated by $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j)$. Intuitively, calibrating is equivalent to replacing forecast $x$ by $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j = x)$ for all possible values $x \in \text{supp}(\tilde{X}_j)$. Perhaps, however, one does not want to work under this particular model. To accommodate alternative models (such as the Gaussian model described in Section 5.2.2.2), the next proposition shows how $\mathbb{Q}$-calibrated forecasts can be transformed into forecasts that are calibrated under some other probability measure $\mathbb{P}$. All the proofs are deferred to Appendix A.

**Proposition 5.2.1.** *Consider a probability measure $\mathbb{P}$ such that $\mathbb{P} \ll \mathbb{Q}$. Let $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denote the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. The forecasts under the new model $\mathbb{P}$ are then given by the transformation $\mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j) = \mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}Y\middle|\mathcal{F}_j\right) / \mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\middle|\mathcal{F}_j\right)$, where $\mathcal{F}_j = \sigma(\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j))$.*

This shows that uncalibrated forecasts from "non-experts" can be calibrated as long as one agrees on some joint distribution for the target outcome and its forecasts. While such constructs certainly deserve further analysis, they are not in the scope of this paper and hence are left for future work. Therefore, from now on, the forecasts are assumed to be calibrated. Note, however, that in general the forecasts should satisfy some minimal performance criterion; simply aggregating entirely arbitrary forecasts is hardly going to lead to improved forecasting accuracy. To this end, Foster and Vohra (1998) analyze probability forecasts and state that "calibration does seem to be an appealing minimal property that any probability forecast should satisfy." They show that one needs to know almost nothing about the outcomes in order to be calibrated. Thus, in theory, calibration can be achieved very easily and overall seems like an appropriate base assumption for developing a general theory of forecast aggregation.

Given that the partial information framework generates all forecast variation from information diversity, it is important to understand the extent to which the forecasters' partial information sets can be measured in practice. First, note that, for the purposes of aggregation, any available information discarded by a forecaster may as well not exist because information comes to the aggregator only through the forecasts. Therefore it is not in any way restrictive to assume that $\mathcal{F}_j = \sigma(X_j)$. Second, the following proposition describes observable measures for the amount of information in each forecast and for the amount of information overlap between any two forecasts.

**Proposition 5.2.2.** *If $\mathcal{F}_j = \sigma(X_j)$ such that $\mathbb{E}(Y|\mathcal{F}_j) = \mathbb{E}(Y|X_j) = X_j$ for all $j = 1, \ldots, N$, then the following holds.*

*i) Forecasts are marginally consistent:* $\mathbb{E}(Y) = \mathbb{E}(X_j)$.

*ii) Variance increases in information:* $\mathrm{Var}\,(X_i) \leq \mathrm{Var}\,(X_j)$ *if* $\mathcal{F}_i \subseteq \mathcal{F}_j$. *Given that* $Y = \mathbb{E}(Y|\mathcal{F})$*, the variances of the forecasts are upper bounded as* $\mathrm{Var}\,(X_j) \leq \mathrm{Var}\,(Y)$ *for all* $j = 1, \ldots, N$.

*iii)* $\mathrm{Cov}\,(X_j, X_i) = \mathrm{Var}\,(X_i)$ *if* $\mathcal{F}_i \subseteq \mathcal{F}_j$. *Again, expressing* $Y = \mathbb{E}(Y|\mathcal{F})$ *implies that* $\mathrm{Cov}\,(X_j, Y) = \mathrm{Var}\,(X_j)$ *for all* $j = 1, \ldots, N$.

This proposition is important for multiple reasons. First, item i) provides guidance in estimating the prior mean of $Y$ from the observed forecasts. Second, item ii) shows that $\mathrm{Var}\,(X_j)$ quantifies the amount of information used by forecaster $j$. In particular, $\mathrm{Var}\,(X_j)$ increases to $\mathrm{Var}\,(Y)$ as forecaster $j$ learns and becomes more informed. Therefore increased variance reflects more information and is deemed helpful. This is a clear contrast to the standard statistical models that often regard higher variance as increased noise and hence harmful. The covariance $\mathrm{Cov}\,(X_i, X_j)$, on the other hand, can be interpreted as the amount of information overlap between forecasters $i$ and $j$. Given that being non-negatively correlated is not generally transitive (Langford et al., 2001), these covariances are not necessarily non-negative even though all forecasts are non-negatively correlated with the outcome. Such negatively correlated forecasts can arise in a real-world setting. For instance, consider two forecasters who see voting preferences of two different sub-populations that are politically opposed to each other. Each individually is a weak predictor of the total vote on any given issue, but they are negatively correlated because of the likelihood that these two blocks will largely oppose each other.

Third and finally, item iii) shows that the covariance matrix $\mathbf{\Sigma}_X$ of the $X_j$s extends to the unknown $Y$ as follows:

$$\mathrm{Cov}\ \left((Y, X_1, \ldots, X_N)'\right) = \begin{pmatrix} \mathrm{Var}\,(Y) & \mathrm{diag}(\mathbf{\Sigma}_X)' \\ \mathrm{diag}(\mathbf{\Sigma}_X) & \mathbf{\Sigma}_X \end{pmatrix}, \tag{5.1}$$

where diag($\mathbf{\Sigma}_X$) denotes the diagonal of $\mathbf{\Sigma}_X$. This is the key to regressing $Y$ on the $X_j$s without a separate training set of past forecasts of known outcomes. The resulting estimator, called the *revealed aggregator*, is

$$X'' := \mathbb{E}(Y|X_1, \ldots, X_N) = \mathbb{E}\left(Y \mid \mathcal{F}''\right),$$

where $\mathcal{F}'' := \sigma(X_1, \ldots, X_N)$ is the $\sigma$-field generated (or information revealed) by the $X_j$s. The revealed aggregator uses all the information that is available in the forecasts and hence is the optimal aggregator under the distribution of $(Y, X_1, \ldots, X_N)$. To make this precise, consider a scoring rule $S(x, y)$ that represents the loss of predicting $x$ when the outcome is $y$. A scoring rule is said to be consistent for the mean of $Y$ if $\mathbb{E}_Y[S(\mathbb{E}_Y(Y), Y)] \leq \mathbb{E}_Y[S(x, Y)]$ for all $x \in \mathbb{R}$. Savage (1971) showed, subject to weak regularity conditions, that all such scoring rules can be written in the form

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x), \tag{5.2}$$

where $\phi$ is a convex function with subgradient $\phi'$. An important special case is the quadratic loss $S(x, y) = (x-y)^2$ that arises when $\phi(x) = x^2$. Now, if an aggregator is defined as any random variable $X \in \sigma(X_1, \ldots, X_N)$, then $X''$ is an aggregator that minimizes expectation of any scoring rule $S$ of the form (5.2):

$$\begin{aligned}
\mathbb{E}[S(X, Y)] &= \mathbb{E}_{X_1,\ldots,X_N}\{\mathbb{E}_{Y|X_1,\ldots,X_N}[S(X, Y)]\} \\
&\geq \mathbb{E}_{X_1,\ldots,X_N}\{\mathbb{E}_{Y|X_1,\ldots,X_N}[S(X'', Y)]\} \\
&= \mathbb{E}[S(X'', Y)].
\end{aligned}$$

Ranjan and Gneiting (2010) showed a similar results for probability forecasts. For these reasons, $X''$ is considered the relevant aggregator under each specific instance of the frame-

work. The next section shows how this aggregator can be captured in practice.

#### 5.2.2.2 Gaussian Partial Information Model

Even though the general framework is convenient for theoretical analysis, it is clearly too abstract for practical applications. Fortunately, applying the framework in practice only requires one extra assumption, namely the choice of a parametric family for the distribution of $(Y, X_1, \ldots, X_N)$. One approach is to refer to Proposition 5.2.2 and choose a family that is parametrized in terms of the first two joint moments. This points at the multivariate Gaussian distribution that is a typical starting point in developing statistical methodology and often provides the cleanest entry into the issues at hand.

The Gaussian distribution is also the most common choice for modeling measurement error. This is typically motivated by assuming the terms to represent sums of a large number of independent sources of error. The central limit theorem then gives a natural motivation for the Gaussian distribution. A similar argument can be made under the partial information framework. First, consider some pieces of information. Each piece either has a positive or negative impact and hence respectively either increases or decreases $Y$. The total sum (integral) of these pieces determines the value of $Y$. Each forecaster, however, only observes the sum of some subset of them. Based on this sum, the forecaster makes an estimate of $Y$. If the pieces are independent and have small tails, then the joint distribution of the forecasters' observations will be asymptotically Gaussian. Given that the number of information pieces in a real-world setup is likely to be large, it makes sense to model the forecasters' observations as jointly Gaussian. Of course, other distributions, such as the multivariate $t$-distribution, are possible. At this point, however, such alternative specifications are best left for future work.

The model variables $(Y, X_1, \ldots, X_N)$ can be modeled directly with a Gaussian distribution as long as they are all real-valued. In many applications, however, $Y$ and $X_j$ may

not be supported on the whole real line. For instance, the aforementioned Good Judgment Project collected probability forecasts of binary events. In this case, $X_j \in [0,1]$ and $Y \in \{0,1\}$. Fortunately, different types of outcome-forecast pairs can be easily addressed by borrowing from the theory of generalized linear models (McCullagh et al., 1989) and utilizing a *link function*. The result is a close yet widely applicable specification called the *Gaussian partial information model*. This model begins by introducing $N+1$ *information variables* that follow a multivariate Gaussian distribution with the covariance pattern (5.1):

$$\begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_N \end{pmatrix} \sim \mathcal{N}_{N+1}\left(\mathbf{0}, \begin{pmatrix} 1 & \text{diag}(\boldsymbol{\Sigma})' \\ \text{diag}(\boldsymbol{\Sigma}) & \boldsymbol{\Sigma} \end{pmatrix}\right) := \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \dots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \dots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \dots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \dots & \delta_N \end{array}\right). \tag{5.3}$$

This distribution supports the Gaussian model similarly to the way the ordinary linear regression supports the class of generalized linear models. In particular, the information variables transform into the outcome and forecasts via an application-specific link function $g(\cdot)$; that is, $Y = g(Z_0)$ and $X_j = \mathbb{E}(Y|Z_j) = \mathbb{E}(g(Z_0)|Z_j)$. Given that $Z_0$ fully determines $Y$, it is sufficient for all information that can be known about $Y$. The remaining variables $Z_1, \dots, Z_N$, on the other hand, summarize the forecasters' partial information. To make this more concrete, consider our two real-world applications. For probability forecasts of a binary event a reasonable link function $g(\cdot)$ is the indicator function $\mathbf{1}_A$, where $A = \{Z_0 > t\}$ for some threshold value $t \in \mathbb{R}$. For real-valued $X_j$ and $Y$, on the other hand, a reasonable choice is a linear function $g(Z_0) = \sigma_0 Z_0 + \mu_0$, where $\mu_0$ and $\sigma_0$ are the prior mean and standard deviation of $Y$, respectively. In general, it makes sense to have $g(\cdot)$ map from the real-numbers to the support of $Y$ such that $Y$ has the correct prior $\mathbb{P}(Y)$.

Overall, this model can be considered as a close yet practical specification of the general framework. After all, it only adds on the assumption of Gaussianity. This extra assumption, however, is enough to allow the construction of the revealed aggregator $X'' = \mathbb{E}(Y|Z_1, \ldots, Z_N)$. For $X''$ and also $X_j$ the conditional expectations can be often computed via the following conditional distributions:

$$Z_0|Z_j \sim \mathcal{N}\left(Z_j, 1 - \delta_j\right) \text{ and}$$
$$Z_0|\mathbf{Z} \sim \mathcal{N}\left(\text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}\mathbf{Z}, 1 - \text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}\text{diag}(\mathbf{\Sigma})\right),$$

where $\mathbf{Z} = (Z_1, \ldots, Z_N)'$. For instance, if both $X_j$ and $Y$ are real-valued, then $X_j = \sigma_0 Z_j + \mu_0$ and $X'' = \text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}(\mathbf{X} - \mu_0\mathbf{1}_N) + \mu_0$, where $\mathbf{X} = (X_1, \ldots, X_N)'$. These conditional distributions arise directly from the well-known conditional distributions of the multivariate Gaussian distribution (see, e.g., Ravishanker and Dey 2001).

### 5.2.3 Previous Work on Model-Based Aggregation

#### 5.2.3.1 Interpreted Signal Framework

The *interpreted signal framework* is a behavioral model that assumes different predictions to arise from differing interpretation procedures (Hong and Page, 2009). For example, consider two forecasters who visit a company and predict its future revenue. One forecaster may carefully examine the company's technological status while the other pays closer attention to what the managers say. Even though the forecasters receive and possibly even use the exact same information, they may interpret it differently and hence end up reporting different forecasts. Therefore forecast heterogeneity is assumed to stem from "cognitive diversity".

This is a very reasonable model and hence has been used in various forms to simulate and illustrate theory about expert behavior (see, e.g., Broomell and Budescu 2009; Parunak

et al. 2013). Consequently, previous authors have constructed many highly specialized toy models of interpreted forecasts. For instance, Dawid et al. (1995) construct simple models of two forecasts to support their discussion on coherent forecast aggregation; Ranjan and Gneiting (2010) use one of these models to simulate calibrated forecasts; and Di Bacco et al. (2003) introduce a model for two forecasters whose (interpreted) log-odds predictions follow a joint Gaussian distribution. Unfortunately, their model is very narrow due to its detailed assumptions and extensive computations. Furthermore, it is not clear how the model can be used in practice or extended to $N$ forecasters. All in all, it seems that successful previous applications of the interpreted signal framework have used it as a basis for illustrating theory instead of actually aiming to model real-world forecasts. In this respect, the framework has remained relatively abstract.

Our partial information framework, however, formalizes the intuition behind it, allows quantitative predictions, and provides a flexible construction for modeling many different forecasting setups. Overall, the framework is very general and, in fact, encompasses all the other authors' models mentioned above as different sub-cases. Unlike the Gaussian model, however, these models make many restrictive assumptions in addition to just choosing a parametric family. Even though the general partial information framework, as described in Section 5.2.2, does not allow the forecasters to interpret information differently and hence does not capture all aspects of the interpreted signal framework, personal interpretations can be easily introduced by associating forecaster $j$ with a probability measure $\mathbb{P}_j$ that describes that forecaster's interpretation of information. If $\mathbb{E}_j$ denotes the expectation under $\mathbb{P}_j$, then it is possible that $X_i = \mathbb{E}_i(Y|\mathcal{F}_i) \neq X_j = \mathbb{E}_j(Y|\mathcal{F}_j)$ even if $\mathcal{F}_i = \mathcal{F}_j$. In practice, however, eliciting the details of each $\mathbb{P}_j$ is hardly possible. Therefore, to keep the model tractable, it is convenient to assume a common interpretation $\mathbb{P}_j = \mathbb{P}$ for all $j = 1, \ldots, N$.

#### 5.2.3.2 Measurement Error Framework

In the absence of a quantitative interpreted signal model, prior applications have typically explained forecast heterogeneity with standard statistical models. These models are different formalizations of the *measurement error framework* that generates forecast heterogeneity purely from a probability distribution. More specifically, this framework assumes a "true" (possibly transformed) forecast $\theta$, which can be interpreted as the prediction made by an ideal forecaster. The forecasters then somehow measure $\theta$ with mean-zero idiosyncratic error. For instance, in our probability forecasting application one possible measurement error model is

$$\begin{aligned} Y &\sim \text{Bernoulli}(\theta), \\ \text{logit}(X_j) &= \text{logit}(\theta) + e_j, \text{ and} \\ e_j &\overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \text{ for all } j = 1, \ldots, N, \end{aligned} \tag{5.4}$$

where $\text{logit}(x) = \log(x/(1-x))$ is the log-odds operator. Given that the errors are generally assumed to have mean zero, measurement error forecasts are unbiased estimates of $\theta$, that is, $\mathbb{E}(X_j|\theta) = \theta$. Observe that this is not the same as assuming calibration $\mathbb{E}(Y|X_j) = X_j$. Therefore an unbiased estimation model is very different from a calibrated model. This distinction is further emphasized by the fact that $X''$ never reduces to a (non-trivial) weighted average of the forecasts (Satopää and Ungar, 2015). Given that the measurement-error aggregators are often different types of weighted averages, measurement error and information diversity are not only philosophically different but they also require very different aggregators.

Example (5.4) illustrates the main advantages of the measurement error framework: simplicity and familiarity. Unfortunately, there are a number of disadvantages. First, measurement-error aggregators estimate $\theta$ instead of the realized value of the random vari-

able $Y$. For this reason, these aggregators often do not satisfy even the minimal performance requirements. For instance, a non-trivial weighted average of calibrated forecasts is necessarily both uncalibrated and under-confident (Ranjan and Gneiting, 2010; Satopää and Ungar, 2015). Second, the standard assumption of conditional independence of the observations forces a specific and highly unrealistic structure on interpreted forecasts (Hong and Page, 2009). Measurement-error aggregators also cannot leave the convex hull of the individual forecasts, which further contradicts the interpreted signal framework (Parunak et al., 2013) and can be easily seen to result in poor empirical performance on many datasets. Third, the underlying model is rather implausible. Relying on a true forecast $\theta$ invites philosophical debate, and even if one assumes the existence of such a value, it is difficult to believe that the forecasters are actually seeing it with independent noise. Therefore, whereas the interpreted signal framework proposes a plausible micro-level explanation, the measurement error model does not; at best, it forces us to imagine a group of forecasters who apply the same procedures to the same data but with numerous small mistakes.

## 5.3 Model Estimation

This section describes methodology for estimating the *information structure* $\mathbf{\Sigma}$. Even though $\mathbf{\Sigma}$ is mostly used for aggregation, it also describes the information among the forecasters (see end of Section 5.2.2.1) and hence should be of interest to decision analysts, psychologists, and the broader community studying collective problem solving. Unfortunately, estimating $\mathbf{\Sigma}$ in full generality based on a single prediction per forecaster is difficult. Therefore, to facilitate model estimation, the forecasters are assumed to predict $K \geq 2$ related events. For instance, in our second application $416$ undergraduates guessed the weights of $20$ people. This yielded a $20 \times 416$ matrix that was then used to estimate $\mathbf{\Sigma}$.

### 5.3.1 General Estimation Problem

Denote the outcome of the $k$th event with $Y_k$ and the $j$th forecaster's prediction for this outcome with $X_{jk}$. For the sake of generality, this section does not assume any particular link function but instead operates directly with the corresponding information variables, denoted with $Z_{jk}$. In practice, the forecasts $X_{jk}$ can be often transformed into $Z_{jk}$ at least approximately. This is illustrated in Section 5.4. Recall that aggregation cannot access to the outcomes $\{Y_1, \ldots, Y_K\}$ or their corresponding information variables $\{Z_{01}, \ldots, Z_{0K}\}$. Instead, $\mathbf{\Sigma}$ is estimated only based on $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_K\}$, where the vector $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{Nk})'$ collects the forecasters' information about the $k$th event.

This estimation must respect the covariance pattern (7.3). More specifically, if $\mathcal{S}_+^N$ denotes the set of $N \times N$ symmetric positive semidefinite matrices and

$$h(\mathbf{M}) := \begin{pmatrix} 1 & \text{diag}(\mathbf{M})' \\ \text{diag}(\mathbf{M}) & \mathbf{M} \end{pmatrix}$$

for some symmetric matrix $\mathbf{M}$, then the final estimate must satisfy the condition $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$. Intuitively, this is satisfied if there exists a random variable $Y$ for which the forecasts $X_j$ are jointly calibrated. In terms of information, this means that it is physically possible to allocate information about $Y$ among the $N$ forecasters in the manner described by $\mathbf{\Sigma}$. Therefore the condition is named *information coherence*.

Unfortunately, simply finding an accurate estimate of $\mathbf{\Sigma}$ does not guarantee precise aggregation. To see this, recall from Section 5.2.2.2 that $\mathbb{E}(Z_{0k}|\mathbf{Z}_k) = \text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}\mathbf{Z}_k$. This term is generally found in the revealed aggregator and hence deserves careful treatment. Re-express the term as $\mathbf{v}'\mathbf{Z}_k$, where $\mathbf{v}$ is the solution to $\text{diag}(\mathbf{\Sigma}) = \mathbf{\Sigma}\mathbf{v}$. The rate at which the solution changes with respect to a change in $\text{diag}(\mathbf{\Sigma})$ depends on the condition number $\text{cond}(\mathbf{\Sigma}) := \lambda_{max}(\mathbf{\Sigma})/\lambda_{min}(\mathbf{\Sigma})$, i.e., the ratio between the maximum and minimum eigenvalues of $\mathbf{\Sigma}$. If the condition number is very large, a small error in $\text{diag}(\mathbf{\Sigma})$ can cause a

large error in $\mathbf{v}$. If the condition number is small, $\boldsymbol{\Sigma}$ is called *well-conditioned* and error in $\mathbf{v}$ will not be much larger than the error in $\text{diag}(\boldsymbol{\Sigma})$. Thus, to prevent estimation error from being amplified during aggregation, the estimation procedure should require $\text{cond}(\boldsymbol{\Sigma}) \leq \kappa$ for a given threshold $\kappa \geq 1$.

This all gives the following general estimation problem:

$$\begin{aligned} &\text{minimize } f_0\left(\boldsymbol{\Sigma}, \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}\right) \\ &\text{subject to } h(\boldsymbol{\Sigma}) \in \mathcal{S}_+^{N+1}, \text{ and} \\ &\qquad \text{cond}(\boldsymbol{\Sigma}) \leq \kappa, \end{aligned} \tag{5.5}$$

where $f_0$ is some objective function. The feasible region defined by the two constraints is convex. Therefore, if $f_0$ is convex in $\boldsymbol{\Sigma}$, expression (5.5) is a convex optimization problem. Typically the global optimum to such a problem can be found very efficiently. Problem (5.5), however, involves $\binom{N+1}{2}$ variables. Therefore it can be solved efficiently with standard optimization techniques, such as the interior point methods, as long as the number of variables is not too large, say, not more than 1,000. Unfortunately, this means that the procedure cannot be applied to prediction polls with more than about $N = 45$ forecasters. This is very limiting as many prediction polls involve hundreds of forecasters. For instance, our two real-world applications involve $100$ and $416$ forecasters. Fortunately, by choosing the loss function carefully one can perform dimension reduction and estimate $\boldsymbol{\Sigma}$ under a much larger $N$. This is illustrated in the following subsections.

### 5.3.2 Maximum Likelihood Estimator

Under the Gaussian model the information structure $\boldsymbol{\Sigma}$ is a parameter of an explicit likelihood. Therefore estimation naturally begins with the maximum likelihood approach (MLE). Unfortunately, the Gaussian likelihood is not convex in $\boldsymbol{\Sigma}$. Consequently, only a locally optimal solution is guaranteed with standard optimization techniques. Further-

more, it is not clear whether the dimension of this form can be reduced. Won and Kim (2006) discuss the MLE under a condition number constraint. They are able to transform the original problem with $\binom{N+1}{2}$ variables to an equivalent problem with only $N$ variables, namely the eigenvalues of $\mathbf{\Sigma}$. This transformation, however, requires an orthogonally invariant problem. Given that the constraint $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$ is not orthogonally invariant, the same dimension-reduction technique cannot be applied. Instead, the MLE must be computed with the $\binom{N+1}{2}$ variables, making estimation slow for small $N$ and undoable even for moderately large $N$. For these reasons the MLE is not discussed further in this paper.

### 5.3.3 Least Squares Estimator

Past literature has discussed many simple covariance estimators that can be applied efficiently to large amounts of data. Unfortunately, these estimators are not guaranteed to satisfy the conditions in (5.5). This section introduces a correctional procedure that inputs any covariance estimator $\mathbf{S}$ and modifies it minimally such that the end result satisfies the conditions in (5.5). More specifically, $\mathbf{S}$ is projected onto the feasible region. This approach, sometimes known as the least squares approach (LSE), motivates a convex loss function that guarantees a globally optimal solution and facilitates dimension reduction. Most importantly, however, it provides a general tool for estimating $\mathbf{\Sigma}$, regardless whether one is working with a Gaussian model or possibly some future non-Gaussian model.

From the computational perspective, it is more convenient to project $h(\mathbf{S})$ instead of $\mathbf{S}$. Even though this could be done under many different norms, for the sake of simplicity, this paper only considers the squared Frobenius norm $||\mathbf{M}||_F^2 = \text{tr}(\mathbf{M}'\mathbf{M})$, where $\text{tr}(\cdot)$ is the trace operator. The LSE is then given by $h^{-1}(\mathbf{\Omega})$, i.e., $\mathbf{\Omega}$ without the first row and column,

where $\mathbf{\Omega}$ is the solution to

$$\begin{aligned} &\text{minimize } ||\mathbf{\Omega} - h(\boldsymbol{S})||_F^2 \\ &\text{subject to } \mathbf{\Omega} \in \mathcal{S}_+^{N+1}, \\ &\qquad \text{cond}(\mathbf{\Omega}) \leq \kappa, \text{ and} \\ &\qquad \text{tr}(\mathbf{A}_j\mathbf{\Omega}) = b_j, \qquad (j = 1, \ldots, N+1). \end{aligned} \tag{5.6}$$

Both $\mathbf{A}_j$ and $b_j$ are constants defined to maintain the covariance pattern (7.3). More specifically, if $\mathbf{e}_j$ denotes the $j$th standard basis vector of length $N+1$, then

$$\begin{aligned} &b_1 = 1, \mathbf{A}_1 = \mathbf{e}_1\mathbf{e}_1', \text{ and} \\ &b_j = 0, \mathbf{A}_j = \mathbf{e}_j\mathbf{e}_j' - 0.5(\mathbf{e}_1\mathbf{e}_j' + \mathbf{e}_j\mathbf{e}_1') \text{ for } j = 2, \ldots, N+1. \end{aligned}$$

If $\mathbf{\Omega}$ satisfies the other two conditions, namely $\mathbf{\Omega} \in \mathcal{S}_+^{N+1}$ and $\text{cond}(\mathbf{\Omega}) \leq \kappa$, then $\mathbf{\Sigma} = h^{-1}(\mathbf{\Omega})$ also satisfies them. This follows from the fact that $\mathbf{\Sigma}$ is a principal sub-matrix of $\mathbf{\Omega}$. Therefore $\mathbf{\Omega} \in \mathcal{S}_+^{N+1}$ implies $\mathbf{\Sigma} \in \mathcal{S}_+^{N}$. Furthermore, Cauchy's interlace theorem (see, e.g., Hwang 2004) states that $\lambda_{min}(\mathbf{\Omega}) \leq \lambda_{min}(\mathbf{\Sigma})$ and $\lambda_{max}(\mathbf{\Sigma}) \leq \lambda_{max}(\mathbf{\Omega})$ such that $\text{cond}(\mathbf{\Sigma}) \leq \text{cond}(\mathbf{\Omega}) \leq \kappa$. Of course, requiring $\text{cond}(\mathbf{\Omega}) \leq \kappa$ instead of $\text{cond}(\mathbf{\Sigma}) \leq \kappa$ shrinks the region of feasible $\mathbf{\Sigma}$s. At this point, however, the exact value of $\kappa$ is arbitrary and merely serves to control $\text{cond}(\mathbf{\Sigma})$. Section 5.3.4 introduces a procedure for choosing $\kappa$ from the data. Under such an adaptive procedure, problem (5.6) can be considered equivalent to directly projecting $\mathbf{S}$ onto the feasible region.

The first step towards solving (5.6) is to express the feasible region as an intersection of the following two sets:

$$\begin{aligned} \mathcal{C}_{sd} &= \left\{\mathbf{\Omega} : \mathbf{\Omega} \in \mathcal{S}_+^{N+1}, \text{cond}(\mathbf{\Omega}) \leq \kappa\right\}, \text{ and} \\ \mathcal{C}_{lin} &= \left\{\mathbf{\Omega} : \text{tr}(\mathbf{A}_j\mathbf{\Omega}) = b_j, j = 1, \ldots, N+1\right\}. \end{aligned}$$

Given that both of these sets are convex, projecting onto their intersection can be computed with the Directional Alternating Projection Algorithm (Gubin et al., 1967). This method makes progress by repeatedly projecting onto the sets $\mathcal{C}_{sd}$ and $\mathcal{C}_{lin}$. Consequently, it is efficient only if projecting onto each of the individual sets is fast. Fortunately, as will be shown next, this turns out to be the case.

First, projecting an $(N+1) \times (N+1)$ symmetric matrix $\mathbf{M} = \{m_{ij}\}$ onto $\mathcal{C}_{lin}$ is a linear map. To make this more specific, let $\mathbf{m} = \text{vec}(\mathbf{M})$ be a column-wise vectorization of $\mathbf{M}$. If $\mathbf{A}$ is a matrix with the $j$th row equal to $\text{vec}(\mathbf{A}_j)$, the linear constraints in (5.6) can be expressed as $\mathbf{Am} = \boldsymbol{e}_1$. Then, the projection of $\mathbf{M}$ onto $\mathcal{C}_{lin}$ is given by $\text{vec}^{-1}(\mathbf{m} + \mathbf{A}'(\mathbf{AA}')^{-1}(\boldsymbol{e}_1 - \mathbf{Am}))$. This expression simplifies significantly by close inspection. In fact, it is equivalent to setting $m_{11} = 1$ and for $j \geq 2$ replacing $m_{j1}$, $m_{1j}$, and $m_{jj}$ by their average $(m_{jj} + m_{j1} + m_{1j})/3$. Denote this projection with the operator $\mathcal{P}_{lin}(\cdot)$.

Second, Tanaka and Nakata (2014) describe a univariate optimization problem that is almost equivalent to projecting $\mathbf{M}$ onto $\mathcal{C}_{sd}$. The only difference is that their solution set also includes the zero-matrix $\mathbf{0}$. Assuming that such a limiting case can be safely handled in the implementation, their approach offers a fast projection onto $\mathcal{C}_{sd}$ even for a moderately large $N$. To describe this approach, consider the spectral decomposition $\mathbf{M} = \mathbf{Q}\text{Diag}(l_1, \ldots, l_{N+1})\mathbf{Q}'$ and the univariate function

$$\pi(\mu) = \sum_{i=1}^{N+1} \left[ (\mu - l_i)_+^2 + (l_i - \kappa\mu)_+^2 \right],$$

where $\text{Diag}(\mathbf{x})$ is a diagonal matrix with diagonal $\mathbf{x}$ and $(\cdot)_+$ is the positive part operator. The function $\pi(\mu)$ can be minimized very efficiently by solving a series of smaller convex problems, each with a closed form solution. The result is a binary-search-like procedure

described by Algorithm 3 in Appendix A. If $\mu^* = \arg\min_{\mu \geq 0} \pi(\mu)$ and

$$\lambda_j^* := \begin{cases} \mu^* & \text{if } l_j \leq \mu^* \\ \kappa\mu^* & \text{if } \kappa\mu^* \leq l_j \\ l_j & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, N+1$, then $\mathbf{Q}\text{Diag}(\lambda_1^*, \ldots, \lambda_{N+1}^*)\mathbf{Q}$ is the projection of $\mathbf{M}$ onto $\mathcal{C}_{sd}$. Call this projection $\mathcal{P}_{sd}(\cdot : \kappa)$.

Algorithm 1 uses these projections to solve (5.6). Each iteration projects twice on one set and once on the other set. The general form of the algorithm does not specify which projection should be called twice. Therefore, given that $\mathcal{P}_{sd}(\cdot : \kappa)$ takes longer to run than $\mathcal{P}_{lin}(\cdot)$, it is beneficial to choose to call $\mathcal{P}_{lin}(\cdot)$ twice. The complexity of each iteration is determined largely by the spectral decomposition which is fairly fast for moderately large $N$. Overall time to convergence, of course, depends on the choice of the stopping criterion. Many intuitive criteria are possible. Given that $\mathbf{\Omega}_D \in \mathcal{C}_{lin}$ and $\mathbf{\Omega}_C \in \mathcal{C}_{sd}$, the stopping criterion $\max\{(\mathbf{\Omega}_D - \mathbf{\Omega}_C)_{ij}^2\} < \epsilon$ suggests that the return value is in $\mathcal{C}_{sd}$ and close to $\mathcal{C}_{lin}$ in every direction. Based on our experience, the algorithm converges quite quickly. For instance, our implementation in C++ generally solves (5.6) for $\epsilon = 10^{-5}$ and $N = 100$ in less than a second on a 1.7 GHz Intel Core i5 computer. This code will be made available online upon publication. For the remainder of the paper, projecting $\mathbf{S}$ onto the feasible region is denoted with the operator $\mathcal{P}_{LSE}(\mathbf{S} : \kappa)$.

### 5.3.4 Selecting $\kappa$

The estimation procedure described in the previous section has one tuning parameter, namely the condition number threshold $\kappa$. This subsection discusses an in-sample approach, called *conditional validation*, that can be used for choosing any tuning parameter,

**Algorithm 1** This procedure projects $h(\boldsymbol{S})$ onto the intersection $\mathcal{C}_{sd} \cap \mathcal{C}_{lin}$. Denote the projection with $\mathcal{P}_{LSE}(\mathbf{S} : \kappa)$. Throughout the paper, the stopping criterion is fixed at $\epsilon = 10^{-5}$.

**Require:** Unconstrained covariance matrix estimator $\mathbf{S}$, stopping criterion $\epsilon > 0$, and an upper bound on the condition number $\kappa \geq 1$.

**procedure** DIRECTIONAL ALTERNATING PROJECTION ALGORITHM
  $\boldsymbol{\Omega}_A \leftarrow h(\boldsymbol{S})$
  **repeat**
    $\boldsymbol{\Omega}_B \leftarrow \mathcal{P}_{lin}(\boldsymbol{\Omega}_A)$
    $\boldsymbol{\Omega}_C \leftarrow \mathcal{P}_{sd}(\boldsymbol{\Omega}_B : \kappa)$
    $\boldsymbol{\Omega}_D \leftarrow \mathcal{P}_{lin}(\boldsymbol{\Omega}_C)$
    $\Delta \leftarrow ||\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_C||_F^2 / \text{tr}\left[(\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_D)'(\boldsymbol{\Omega}_B - \boldsymbol{\Omega}_C)\right]$
    $\boldsymbol{\Omega}_A \leftarrow \boldsymbol{\Omega}_B + \Delta(\boldsymbol{\Omega}_D - \boldsymbol{\Omega}_B)$
  **until** $\max\left\{(\boldsymbol{\Omega}_D - \boldsymbol{\Omega}_C)_{ij}^2\right\} < \epsilon$
  **return** $h^{-1}(\boldsymbol{\Omega}_C)$
**end procedure**

such as $\kappa$, under the partial information framework. To motivate, recall that the revealed aggregator $X''$ uses $\boldsymbol{\Sigma}$ to regress $Z_0$ on the rest of the $Z_j$s. Of course, the accuracy of this prediction cannot be known until the actual outcome is observed. However, apart from being unobserved, the variable $Z_0$ is theoretically no different to the other $Z_j$s. This suggests the following algorithm: for some value $\nu$ compute $\mathcal{P}_{LSE}(\mathbf{S} : \nu)$, let each of the $Z_j$s in turn play the role of $Z_0$, predict its value based on $Z_i$ for $i \neq j$, and choose the value of $\nu$ that yields the best overall accuracy. Even though many accuracy measures could be chosen, this paper uses the conditional log-likelihood. Therefore, if $\mathbf{Z}_j^* = (Z_{j1}, \ldots, Z_{jK})'$ collects the $j$th forecaster's information about the $K$ events, the chosen value of $\kappa$ is

$$\kappa_{cov} = \underset{\nu \geq 1}{\arg\max} \sum_{j=1}^{N} \ell\left(\mathbf{Z}_j^*, \mathcal{P}_{LSE}(\mathbf{S} : \nu) \middle| \mathbf{Z}_i^* \text{ for } i \neq j\right), \tag{5.7}$$

where the log-likelihood is now conditional on $\mathbf{Z}_i^*$s for $i \neq j$ and $\mathbf{S}$ is computed based on all the forecasts $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_N^*$. Plugging this into the projection algorithm gives the final estimate $\boldsymbol{\Sigma}_{cov} := \mathcal{P}_{LSE}(\mathbf{S} : \kappa_{cov})$.

Unfortunately, the optimization problem (5.7) is non-convex in $\nu$. However, as was

mentioned before, Algorithm 1 is fast for moderately sized $N$. Therefore $\kappa$ can be chosen efficiently (possibly in parallel on multicore machines) over a grid of candidate values. Overall, the idea in conditional validation is similar to cross-validation but, instead of predicting across rows (observations), the prediction is performed across columns (variables). This not only mimics the actual process of revealed aggregation but is also likely to be more appropriate for prediction polling that typically involves a large number of forecasters (large $N$) predicting relatively few events (small $K$). Furthermore, it has no tuning parameters and remains more stable when $K$ is small; see Appendix B for an illustration of this result under synthetic data.

## 5.4 Applications

This section applies the partial information framework to different types of real world forecasts. For each type there may be different ways to adopt the Gaussian model. The main point, however, is not to find the optimal way to do this but rather to give illustrative examples on using the framework and also to show how the resulting partial information aggregators outperform the commonly used measurement error aggregators.

### 5.4.1 Probability Forecasts of Binary Outcomes

#### 5.4.1.1 Dataset

During the second year of the Good Judgment Project (GJP) the forecasters made probability estimates for 78 events, each with two possible outcomes. One of these events was illustrated in Figure 5.1. Each prediction problem had a timeframe, defined as the number of days between the first day of forecasting and the anticipated resolution day. These timeframes varied largely among problems, ranging from 12 days to 519 days with a mean of 185.4 days. During each timeframe the forecasters were allowed to update their predictions

as frequently as they liked. The forecasters knew that their estimates would be assessed for accuracy using the quadratic loss (often known as the Brier score; see Brier 1950 for more details). This is a proper loss function that incentivized the forecasters to report their true beliefs instead of attempting to game the system. In addition to receiving \$150 for meeting minimum participation requirements that did not depend on prediction accuracy, the forecasters received status rewards for their performance via leader-boards displaying the losses for the best $20$ forecasters. Depending on the details of the reward structure, such a competition for rank may eliminate the truth-revelation property of proper loss functions (see, e.g., Lichtendahl Jr and Winkler 2007).

This data collection raises several issues. First, given that the current paper does not focus on modeling dynamic data, only forecasts made within some common time interval should be considered. Second, not all forecasters made predictions for all the events. Furthermore, the forecasters generally updated their forecasts infrequently, resulting into a very sparse dataset. Such high sparsity can cause problems in computing the initial unconstrained estimator $\mathbf{S}$. Evaluating different techniques to handle missing values, however, is well outside the scope of this paper. Therefore, to somewhat alleviate the effect of missing values, only the hundred most active forecasters are considered. This makes sufficient overlap highly likely but, unfortunately, still not guaranteed.

All these considerations lead to a parallel analysis of three scenarios: High Uncertainty (HU), Medium Uncertainty (MU), and Low Uncertainty (LU). Important differences are summarized in Table 5.1. Each scenario considers the forecasters' most recent prediction within a different time interval. For instance, LU only includes each forecaster's most recent forecast during $30 - 60$ days before the anticipated resolution day. The resulting dataset has $60$ events of which $13$ occurred. In the corresponding $60 \times 100$ table of forecasts, around 42 % of the values are missing. The other two scenarios are defined similarly.

Table 5.1: Summary of the three time intervals analyzed. Each scenario considers the forecasters' most recent forecasts within the given time interval. The value in the parentheses represent the number of events occurred. The final column shows the percent of missing forecasts.

| Scenario | Time Interval | # of Events | Missing (%) |
|---|---|---|---|
| High Uncertainty (HU) | $90-120$ | $49\ (10)$ | $51$ |
| Medium Uncertainty (MU) | $60-90$ | $56\ (14)$ | $46$ |
| Low Uncertainty (LU) | $30-60$ | $60\ (13)$ | $42$ |

#### 5.4.1.2 Model Specification and Aggregation

The first step is to pick a link function and derive a Gaussian model for probability forecasts of binary events. Overall, this construction resembles in many ways the latent variable version of a standard probit model.

**Model Instance.** Identify the $k$th event with $Y_k \in \{0, 1\}$. These outcomes link to the information variables via the following function:

$$Y_k = g(Z_{0k}) = \begin{cases} 1 & \text{if } Z_{0k} > t_k \\ 0 & \text{otherwise,} \end{cases}$$

where $t_k \in \mathbb{R}$ is some threshold value. Therefore the link function $g(\cdot)$ is simply the indicator function $\mathbf{1}_{A_k}$ of the event $A_k = \{Z_{0k} > t_k\}$. This threshold is defined by the prior probability of the $k$th event $\mathbb{P}(Y_k = 1) = \Phi(-t_k)$, where $\Phi(\cdot)$ is the CDF of a standard Gaussian distribution. Given that the thresholds are allowed to vary among the events, each event has its own prior. The corresponding probability forecasts $X_{jk} \in [0, 1]$ are

$$X_{jk} = \mathbb{E}(Y_k | Z_{jk}) = \Phi\left(\frac{Z_{jk} - t_k}{\sqrt{1 - \delta_j}}\right).$$

In a similar manner, the revealed aggregator $X_k'' \in [0,1]$ for event $k$ is

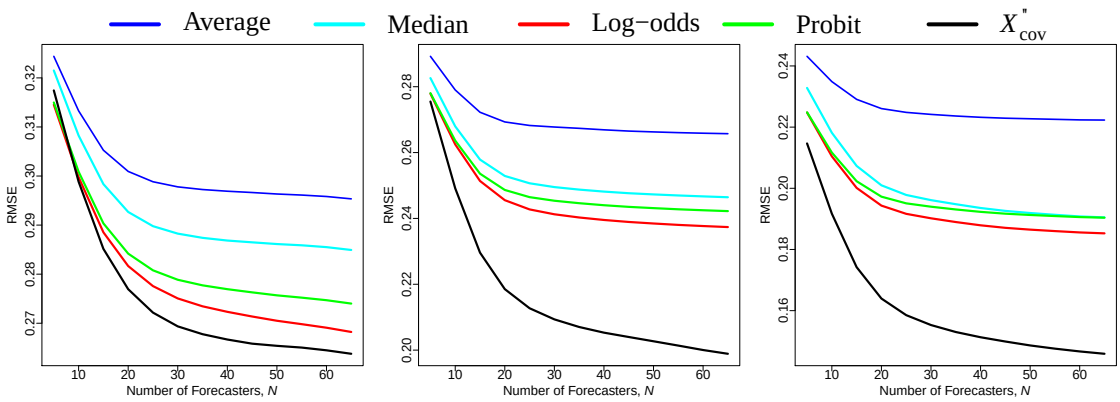
$$X_k'' = \mathbb{E}(Y_k|\mathbf{Z}_k) = \Phi\left(\frac{\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_k - t_k}{\sqrt{1-\text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\text{diag}(\boldsymbol{\Sigma})}}\right). \tag{5.8}$$

All the parameters of this model can be estimated from the data. The first step is to specify a version of the unconstrained estimate $\mathbf{S}$. If the $t_k$'s do not change much, a reasonable and simple estimate is obtained by transforming the sample covariance matrix $\mathbf{S}_P$ of the probit scores $P_{jk} := \Phi^{-1}(X_{jk})$. More specifically, if $\mathbf{D} := \text{Diag}(\mathbf{d})\text{Diag}(\mathbf{1}+\mathbf{d})^{-1}$, where $\mathbf{d} = \text{diag}(\mathbf{S}_P)$, then an unconstrained estimator of $\boldsymbol{\Sigma}$ is given by $\mathbf{S} = (\mathbf{I}_N - \mathbf{D})^{1/2}\mathbf{S}_P(\mathbf{I}_N - \mathbf{D})^{1/2}$. Recall that the GJP data holds many missing values. This is handled by estimating each pairwise covariance in $\mathbf{S}_P$ based on all the events for which both forecasters made predictions. Next, compute $\boldsymbol{\Sigma}_{cov}$, where $\kappa_{cov}$ is chosen over a grid of $100$ candidate values between $10$ and $1,000$. Finally, the threshold $t_k$ can be estimated by letting $\mathbf{P}_k = (P_{1k},\ldots,P_{Nk})'$, observing that $-\text{Diag}(\mathbf{1}-\text{diag}(\boldsymbol{\Sigma}))^{1/2}\mathbf{P}_k \sim \mathcal{N}_N(t_k\mathbf{1}_N, \boldsymbol{\Sigma})$, and computing the precision-weighted average:

$$\hat{t}_k = -\frac{\mathbf{P}_k'\text{Diag}(\mathbf{1}-\text{diag}(\boldsymbol{\Sigma}_{cov}))^{1/2}\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}}.$$

If $\mathbf{P}_k$ has missing values, the corresponding rows and columns of $\boldsymbol{\Sigma}_{cov}$ are dropped. Intuitively, this estimator gives more weight to the forecasters with very little information. These estimates are then plugged in to (5.8) to get the revealed aggregator $X_{cov}''$.

This aggregator is benchmarked against the state-of-the-art measurement-error aggregators, namely the average probability, median probability, average probit-score, and average log-odds. Unequally weighted averages were not considered because it is unclear how the weights would be determined based on forecasts alone, and even if this could be done somehow (perhaps based on self-assessment or organizational status), using unequal weights often leads to no or very small performance gains (Rowse et al., 1974; Ashton




(a) High Uncertainty (HU) (b) Medium Uncertainty (MU) (c) Low Uncertainty (LU)

Figure 5.3: Average prediction accuracy over the 1,000 sub-samplings of the forecasters. See Table 5.1 for descriptions of the different scenarios.

and Ashton, 1985; Flores and White, 1989). To avoid infinite log-odds and probit scores, extreme forecasts $X_{jk} = 0$ and $1$ were censored to $X_{jk} = 0.001$ and $0.999$, respectively. The results remain insensitive to the exact choice of censoring as long as this is done in a reasonable manner to keep the extreme probabilities from becoming highly influential in the logit- or probit-space. The accuracy of the aggregates is measured with the average root-mean-squared-error (RMSE). Note that this is nothing but the square root of the commonly used Brier score. Instead of considering all the forecasts at once, the aggregators are evaluated under different $N$ via repeated subsampling of the $100$ most active forecasters; that is, choose $N$ forecasters uniformly at random, aggregate their forecasts, and compute the RMSE. This is repeated 1,000 times with $N = 5, 10, \ldots, 65$ forecasters. Due to high computational cost, the simulation was stopped after $N = 65$. In the rare occasion where no pairwise overlap is available between one or more pairs of the selected forecasters, the subsampling is repeated until all pairs have at least one problem in common.

Figure 5.3 shows the average RMSEs under the three scenarios described in Table 5.1. Here a reasonable upper bound is given by $0.5$ as this is the RMSE one would receive by

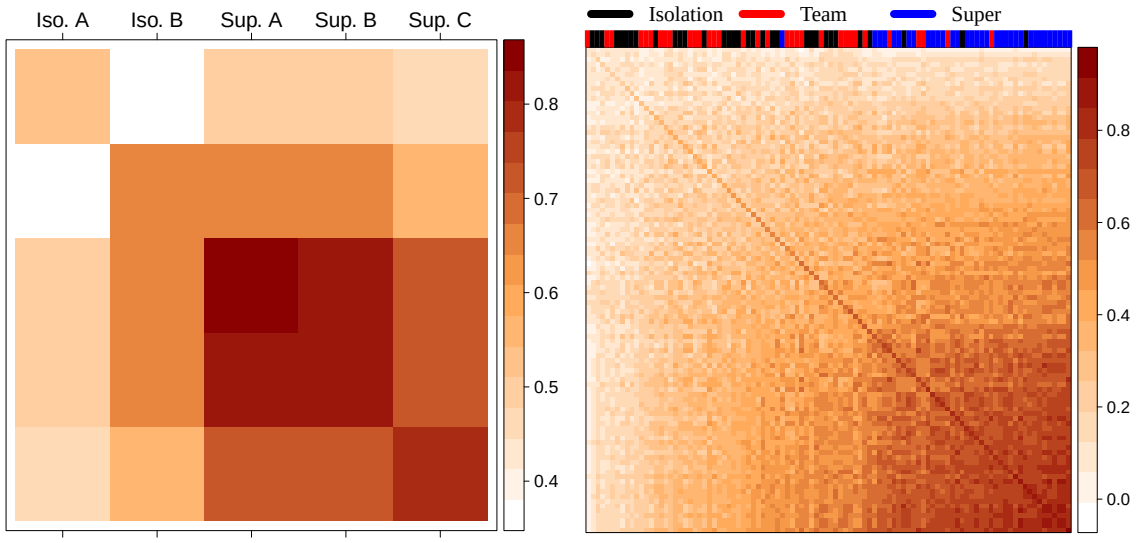
constantly predicting $0.5$. All presented scores, however, are well below it and improve uniformly from left to right, that is, from HU to LU. This reflects the decreasing level of uncertainty. In all the figures the measurement-error aggregators rank in the typical order (from worst to best): average probability, median probability, average probit, and average log-odds. Regardless of the level of uncertainty, the revealed aggregator $X''_{cov}$ outperforms the averaging aggregators as long as $K \geq 10$. The relative advantage, however, increases from HU to LU. More specifically, the improvement from Log-odds to $X''_{cov}$ is about $2\%$, $17\%$, and $21\%$ in HU, MU, and LU, respectively. This trend can be explained by several reasons. First, as can be seen in Table 5.1, the amount of data increases from HU to LU. This yields a better estimate of $\mathbf{\Sigma}$ and hence more accurate revealed aggregation. Second, the forecasters are more likely to be well-calibrated under MU and LU than under HU (see, e.g., Braun and Yaniv 1992). Third, under HU the events are still inherently very uncertain. Consequently, the forecasters are unlikely to hold much useful information as a group. Under such low information diversity, measurement-error aggregators generally perform relatively well (Satopää et al. 2015). In the contrary, under MU the events have lost a part of their inherent uncertainty, allowing some forecasters to possess useful private information. These individuals are then prioritized by $X''_{cov}$ while the averaging-aggregators continue treating all forecasts equally. Consequently, the performance of the measurement error aggregators plateaus after $N = 30$ or so. Therefore having more than about $30$ forecasters does not make a difference if one is determined to aggregate their predictions using the measurement error techniques; a similar results was reported by Satopää et al. 2014. In contrast, however, the RMSE of $X''_{cov}$ continues to improve linearly in $N$, suggesting that $X''_{cov}$ is able to find some residual information in each additional forecaster and use this to increase its performance advantage.

#### 5.4.1.3 Information Diversity

The GJP assigned the forecasters to make predictions either in isolation or in teams. Furthermore, after the first year of the tournament, the top 2% forecasters were elected to the elite group of "super-forecasters." These super-forecasters then worked in exclusive teams to make highly accurate predictions on the same events as the rest of the forecasters. Overall, these assignments directly suggest a level of information overlap. In particular, recalling the interpretation of $\boldsymbol{\Sigma}$ from Section 5.2.2.1, super-forecasters can be expected to have the highest $\delta_j$s and forecasters in the same team should have a relatively high $\rho_{ij}$. This subsection examines how well $\boldsymbol{\Sigma}_{cov}$ aligns with this prior knowledge about the forecasters' information structure.

For the sake of brevity, only the LU scenario is analyzed as this is where $X''_{cov}$ presented the highest relative improvement. The associated 100 forecasters involve 36 individuals predicting in isolation, 33 forecasting team-members (across 24 teams), and 31 super-forecasters (across 5 teams). Figure 5.4a displays $\boldsymbol{\Sigma}_{cov}$ for the five most active forecasters. This group involves two forecasters working in isolation (Iso. A and B) and three super-forecasters (Sup. A, B, and C), of whom the super-forecasters A and B are in the same team. Overall, $\boldsymbol{\Sigma}_{cov}$ agrees with this classification: the only two team members, namely Sup. A and B have a relatively high information overlap. In addition, the three super-forecasters are more informed than the non-super-forecasters. Such a high level of information unavoidably leads to higher information overlap with the rest of the forecasters.

By and large, this agreement generalizes to the entire group of forecasters. To illustrate, Figure 5.4b displays $\boldsymbol{\Sigma}_{cov}$ for all the 100 forecasters. The information structure has been ordered with respect to the diagonal such that the more informed forecasters appear on the right. Furthermore, a colored rug has been appended on the top. This rug shows whether each forecaster worked in isolation, in a non-super-forecaster team, or in a super-forecaster team. Observe that the super-forecasters are mostly situated on the right among the most




(a) $\mathbf{\Sigma}_{cov}$ for the five most active forecasters

(b) $\mathbf{\Sigma}_{cov}$ for all 100 forecasters shows high information diversity.

Figure 5.4: The estimated information structure $\mathbf{\Sigma}$ under the LU scenario. Each forecaster worked either in isolation, in a non-super-forecaster team, or in a super-forecaster team. The super-forecasters generally have more information than the forecasters working in isolation.

informed forecasters. The average estimated $\delta_j$ among the super-forecaster is $0.80$. On the other hand, the average estimated $\delta_j$ among the individuals working in isolation or in non-super-forecaster teams are $0.47$ and $0.50$, respectively. Therefore working in a team makes the forecasters' predictions, on average, slightly more informed.

In general, a plot such as Figure 5.4b is useful for assessing the level of information diversity among the forecasters: the further away it is from a monochromatic plot, the higher the information diversity. That being said, the colorful Figure 5.4b suggests that the GJP forecasters have high information diversity. This makes sense as these forecasters were asked to make predictions about international political events. Given that on such events the

forecasters' background knowledge, education, how closely they follow the news, and so on matter, one should expect a high level of information diversity. Therefore not only does $X''_{cov}$ clearly outperform the common measurement error aggregators in terms of prediction accuracy but the Gaussian model also captures true structure in the data.

### 5.4.2 Point Forecasts of Continuous Outcomes

#### 5.4.2.1 Dataset

Moore and Klein (2008) hired $415$ undergraduates from Carnegie Mellon University to guess the weights of $20$ people based on a series of pictures. These forecasts were illustrated in Figure 5.2. The target people were between 7 and 62 years old and had weights ranging from $61$ to $230$ pounds, with a mean of $157.6$ pounds. All the students were shown the same pictures and hence given the exact same information. Therefore any information diversity arises purely from the participants' decisions to use different subsets of the same information. Consequently, information diversity is likely to be low compared to Section 5.4.1 where diversity also stemmed from differences in the information available to the forecasters.

Unlike in Section 5.4.1, the Gaussian model can be applied almost directly to the data. Only the effect of extreme values was reduced via a $90\%$ Winsorization (Hastings et al., 1947). This handled some obvious outliers. For instance, the original dataset contained a few estimates above $1000$ pounds and as low as $10$ pounds. Winsorization generally improved the performance of all the competing aggregators.

#### 5.4.2.2 Model Specification and Aggregation

**Model Instance.** Suppose $Y_k$ and $X_{jk}$ are real-valued. If the proper non-informative prior distribution of $Y_k$ is $\mathcal{N}(\mu_{0k}, \sigma_0^2)$, then $Y_k = g(Z_{0k}) = Z_{0k}\sigma_0 + \mu_{0k}$. Consequently, $X_{jk} = \mathbb{E}(Y|Z_{jk}) = Z_{jk}\sigma_0 + \mu_{0k}$ for all $j = 1, \ldots, N$. Therefore $X_j \sim \mathcal{N}(\mu_{0k}, \sigma_j^2)$ for some $\sigma_j^2 \leq \sigma_0^2$. If $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{Nk})'$, then the revealed aggregator for the $k$th event is

$$X_k'' = \mathbb{E}\left(Y_k|\mathbf{Z}_k\right) = \text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{Z}_k\sigma_0 + \mu_{0k}. \tag{5.9}$$

Under this model the prior distribution of $Y_k$ is specified by $\mu_{0k}$ and $\sigma_0^2$. Given that $\mathbb{E}(X_{jk}) = \mu_{0k}$ for all $j = 1, \ldots, N$, the sample average $\hat{\mu}_{0k} = \sum_{j=1}^{N} X_{jk}/N$ provides an initial estimate of $\mu_{0k}$. The value of $\sigma_0^2$ can be estimated by assuming a distribution for the $\sigma_j^2$s. More specifically, let $\sigma_j^2$ be i.i.d. on the interval $[0, \sigma_0^2]$ and use the resulting likelihood to estimate $\sigma_0^2$. For instance, a non-informative choice is to assume $\sigma_j^2 \overset{i.i.d.}{\sim} \mathcal{U}(0, \sigma_0^2)$, which leads to the maximum likelihood estimator $\max\{\sigma_j^2\}$. This has a downward bias that can be corrected by a multiplicative factor of $(N+1)/N$. Therefore, replacing $\sigma_j^2$ with the sample variance $s_j = \sum_{k=1}^{K}(X_{jk} - \hat{\mu}_{0k})^2/(K-1)$ gives the final estimate $\hat{\sigma}_0^2 = (N+1)/N \max\{s_j\}$. Using these estimates, the $X_{jk}$s can be transformed into the $Z_{jk}$s whose sample covariance matrix $\mathbf{S}_Z$ provides the unconstrained estimator for the projection algorithm. The value of $\kappa_{cov}$ is chosen over a grid of $10$ values between $10$ and $10,000$. Once $\boldsymbol{\Sigma}_{cov}$ has been computed, the prior means are updated with the precision-weighted averages $\hat{\mu}_{0k} = (\mathbf{X}_k'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}_N)/(\mathbf{1}_N'\boldsymbol{\Sigma}_{cov}^{-1}\mathbf{1}_N)$. In the end, all these estimates are plugged in (5.9) to get the revealed aggregator $X_{cov}''$.

This aggregator is compared against the average, median, and average of the median and average (AMA). The last competitor, namely AMA is a heuristic aggregator that Lobo and Yao (2010) showed to work particularly well on many different real-world forecast-

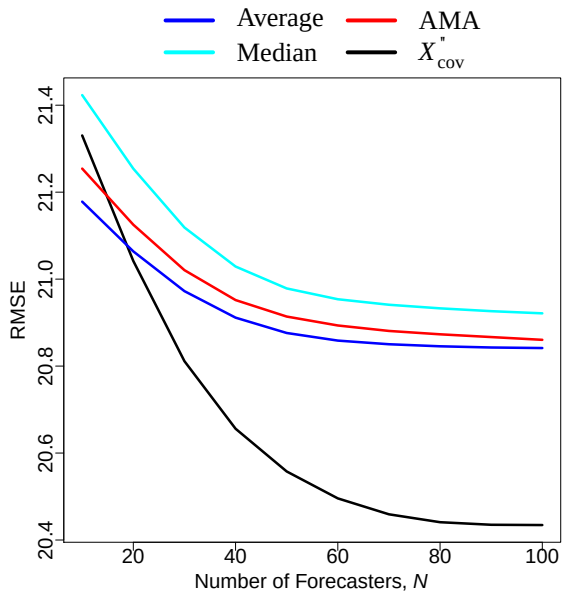



Figure 5.5: Average prediction accuracy

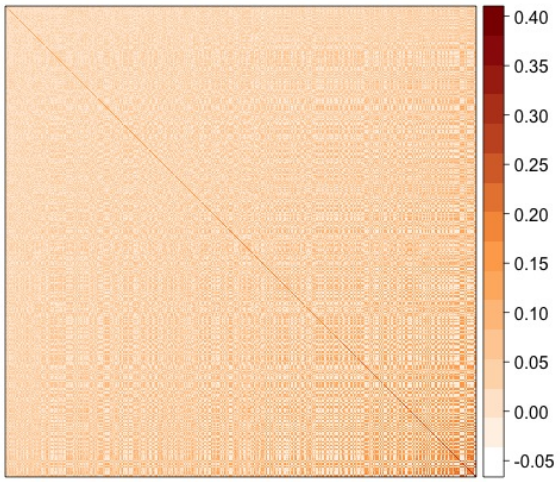



Figure 5.6: $\boldsymbol{\Sigma}_{cov}$ for all 416 forecasters shows low information diversity.

ing datasets. In this section the overall accuracy is measured with the RMSE averaged over $10,000$ sub-samplings of the $416$ participants. That is, each iteration chooses $N$ participants uniformly at random, aggregates their forecasts, and computes the RMSE. The size of the sub-samples is varied between $10$ and $100$ with increments of $10$. These scores are presented in Figure 5.5. The average outperforms the median across all $N$. The performance of AMA falls between that of average and median, reflecting its nature as a compromise of the two. The revealed aggregator $X''_{cov}$ is the most accurate once $N > 10$. The relatively worse performance at $N = 10$ suggests that $10$ observations is not enough to estimate $\hat{\mu}_{0k}$ accurately. As $N$ approaches $100$, however, $X''_{cov}$ collects information efficiently and increases the performance advantage against the other aggregators.

Figure 5.6 shows $\boldsymbol{\Sigma}_{cov}$ for all the 416 forecasters. Similarly to before, the matrix has been ordered such that the most knowledgeable forecasters are on the right. Overall, this plot is much more monochromatic than the one presented earlier in Figure 5.4b, suggesting that information diversity among the 416 students is rather lower. This aligns with the ex-

pectations laid out earlier in Section 5.4.2.1. If there were no information diversity, i.e., all the forecasters used the same information, then averaging aggregators, such as the simple average, would perform very well (Satopää et al., 2015). Such a limiting case, however, is rarely encountered in practice. Often at least some information diversity is present. The results in the current section show that the revealed aggregator does not require extremely high information diversity in order to outperform the measurement-error aggregators.

## 5.5 Discussion

This paper introduced the partial information framework for modeling forecasts from different types of prediction polls. Even though the framework can be used for theoretical analysis and studying information among groups of experts, the main focus was on model-based aggregation of forecasts. Such aggregators do not require a training set. Instead, they operate under a model of forecast heterogeneity and hence can be applied to forecasts alone. Under the partial information framework, all forecast heterogeneity stems from differences in the way the forecasters use information. Intuitively, this is more plausible at the micro-level than the historical measurement error. To facilitate practical applications, the partial information framework motivates and describes the forecasters' information with a patterned covariance matrix (Equation 5.1). A correctional procedure was proposed (Algorithm 1) as a general tool for estimating these information structures. This procedure inputs any covariance estimator and modifies it minimally such that the final output represents a physically feasible allocation of information. Even though the general partial information framework describes an optimal aggregator, it is generally too abstract to be directly applied in practice. As a solution, this paper discusses a close yet practical specification within the framework, known as the Gaussian model (Section 5.2.2.2). The Gaussian model permits a closed-form solution for the optimal aggregator and extends to different types of forecast-outcome pairs via a link function. These partial information aggrega-

tors were evaluated against the common measurement error aggregators on two different real-world (Section 5.4) prediction polls. In each case the Gaussian model outperformed the typical measurement-error-based aggregators, suggesting that information diversity is more important for modeling forecast heterogeneity.

Generally speaking, partial information aggregation works well because it downweights pairs or sets of forecasters that share more information and upweights ones that have unique information (or choose to attend to unique information as is the case, e.g., in Section 5.4.2, where forecasters made judgments based on the same pictures). This is very different from measurement-error aggregators that assume all forecasters to have the same information and hence consider them equally important. While simple measurement-error techniques, such as the average or median, can work well when the forecasters truly operate on the same information set, in real-world prediction polls participants are more likely to have unequal skill and information sets. Therefore prioritizing is almost certainly called for. Of course, the more diverse these sets are, the better the partial information aggregators can be expected to perform relative to the measurement error aggregators. To illustrate this result, compare the relative performances in Section 5.4.1 (high information diversity) against those in Section 5.4.2 (low information diversity).

Overall, the partial information framework can be applied and extended in many different ways. For instance, in this paper the $j$th forecaster's prediction was assumed to be the expectation of $Y$ after observing some partial information $\mathcal{F}_j$. In some applications, however, other constructs, such as the conditional median or other quantiles, may be more appropriate. Such extensions can be handled by considering the distribution of $Y|\mathcal{F}_j$ and then equating the $j$th forecaster's prediction to any desired functional of this distribution. This is particularly easy under the Gaussian model, where $Y|\mathcal{F}_j$ conveniently follows a Gaussian distribution.

In terms of future research, the partial information framework offers both theoretical

and empirical directions. One theoretical avenue involves estimation of information overlap. In some cases the higher order overlaps have been found to be irrelevant to aggregation. For instance, DeGroot and Mortera (1991) show that the pairwise conditional (on the truth) distributions of the forecasts are sufficient for computing the optimal weights of a weighted average. Theoretical results on the significance or insignificance of higher order overlaps under the partial information framework would be desirable. Given that the Gaussian model can only accommodate pairwise information overlap, such a result would reveal the need of a specification that is more complex than the Gaussian model.

A promising empirical direction is the Bayesian approach. These techniques are very natural for fitting hierarchical models such as the ones discussed in this paper. Furthermore, in many applications with small or moderately sized datasets, Bayesian methods have been found to be more stable than the likelihood-based alternatives. Therefore, given that the number of forecasts in a prediction poll is typically quite small, a Bayesian approach is likely to improve the quality of the final aggregate. This would involve developing a prior distribution for the information structure – a problem that seems interesting in itself. Overall, this avenue should certainly be pursued, and the results tested against other high performing aggregators.

## 5.6 Acknowledgments

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or

implied, of IARPA, DoI/NBC, or the U.S. Government. The authors would also like to thank Don Moore for providing us with the weight dataset.

# 6

# Bayesian Aggregation of Two Forecasts in the Partial Information Framework*


## Abstract

The partial information framework was introduced in Satopää et al. (2015, 2016) as a theory for aggregating probability and point forecasts from a pool of expert forecasters. The examples there assume a fixed information overlap model with unknown parameters which are estimated from data. In the present paper, we examine how the framework can be used in a one-shot aggregation problem in which parameters cannot be estimated. Our approach is Bayesian; the proposed estimator is a mixture of the fixed parameter estimators over the posterior distribution of the parameters. We compare this to the aggregator arising from fixed overlap Gaussian models from the partial information framework, as well as to classical aggregators based on other aggregation paradigms.


*Joint work with Philip Ernst, Robin Pemantle, and Lyle H. Ungar

## 6.1 Introduction

The partial information framework for forecast aggregation was introduced by Satopää et al. (2015). This framework in its most general form is a probability model for forecast aggregation, allowing for many possible information structures. A Gaussian model was proposed within this framework. The resulting aggregate forecast depends on parameters which must be estimated. Apparatus to estimate parameters in the Gaussian partial information model was further developed in Satopää et al. (2016).

The purpose of the present note is to show how, in theory, the Gaussian aggregator may be computed via a Bayesian approach that does not require parameter estimation. Our main result is Theorem 6.3.1 below, which gives an explicit formula for the Gaussian aggregator in a one-shot aggregation problem with two forecasters.

In the remainder of the introduction we give a brief description of the problems of event forecasting and forecast aggregation, then summarize the partial information framework, the Gaussian partial information model and our Bayesian approach. Section 6.2 recalls the relevant computations for the Gaussian model with fixed parameters. Section 6.3 computes the Bayesian aggregator. Section 6.5 contains our proposed methodology applied to a real data set on election forecasting for the 2012 presidential election.

### 6.1.1 Event forecasting, loss functions, and calibration

In event forecasting, an expert is asked for a series $\{p_n\}$ of probability forecasts for events $\{A_n\}$. The quantitative study of event forecasting dates back at least three decades (Dawid, 1982; Murphy and Winkler, 1987a). Typically the expert is scored by a loss function $L(p_n, \mathbf{1}_{A_n})$. The loss function $L$ is assumed to be *proper*, meaning that $p$ minimizes $\mathbb{E}L(\cdot, Y)$ when $Y$ is a Bernoulli random variable with mean $p$. Thus a forecaster with subjective probability $p$ minimizes expected loss by forecasting $p$. A more complete discussion of probability forecasting and proper loss functions may be found in Hwang and

Pemantle (1997).

Probability forecasts may suffer from two kinds of error, namely bias and imprecision. Bias occurs when the long run frequency of $A_n$ for those $p_n \approx p$ is not equal to $p$. Imprecision occurs when $p_n$ is typically not close to zero or one. Given a sufficiently long stream of forecasts, these two problems may be separated: each forecast $p_n$ may be replaced by the forecast $q(p_n)$ where $q(t)$ is the long run frequency of $A_n$ given a forecast of $t$. The forecast is then said to be *calibrated* (cf. Murphy and Winkler (1987a)) and the main objective is to minimize loss. In this paper we always assume calibrated forecasts.

### 6.1.2 Forecast aggregation

Forecast aggregation is the problem of producing a synthesized forecast from a collection of expert forecasts. Various probability models have been implicitly or explicitly used for this problem. As discussed in Satopää et al. (2015), if the events are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ then calibration means precisely that an expert's forecast $p$ for an event $A$ is equal to $\mathbb{P}(A|\mathcal{F}')$ for some $\mathcal{F}' \subseteq \mathcal{F}$. The $\sigma$-field $\mathcal{F}'$ represents the information used to make the forecast and is not necessarily the full information available to the expert.

Some empirical work on forecast aggregation operates outside these assumptions. For example, the *measurement error framework* assumes there is a true probability $\theta$ interpreted as the forecast made by an ideal forecaster, and that actual forecasters "observe" some transformation $\phi(\theta)$ together with independent mean zero idiosyncratic errors.

This leads to relatively simple aggregation rules. For example, if $\phi$ is the identity, the forecasters are assumed to be reporting $\theta$ plus independent mean zero errors. The corresponding aggregator simply averages the forecasts:

$$g_{\text{ave}}(p_1, \ldots, p_n) := \frac{1}{n} \sum_{k=1}^{n} p_k \,. \tag{6.1}$$

When $\phi$ is the inverse normal CDF, this leads to *probit averaging*, defined by

$$g_{\text{probit}}(p_1, \ldots, p_n) := \Phi\left(\frac{1}{n}\sum_{k=1}^{n}\Phi^{-1}(p_k)\right). \tag{6.2}$$

Such models, while very common in practice, lead both to uncalibrated forecasts and sub-optimal performance. Theoretical problems with these models are discussed by Hong and Page (2009); for example, such aggregators can never leave the convex hull of the individual expert forecasts, which is demonstrably sub-optimal in some cases (Parunak et al., 2013); see also Satopää et al. (2016, Section 2.3.2).

Satopää et al. (2015), propose the *partial information framework*. This model for aggregation of calibrated forecasts assumes that each forecaster $i$, $1 \leq i \leq N$, has access to information $\mathcal{F}_i$. The aggregator has access only to the forecasts $p_i := \mathbb{P}(A|\mathcal{F}_i)$. The theoretical best forecast with this information is the *revealed estimator*

$$p_* := \mathbb{P}(A|p_i : 1 \leq i \leq N).$$

It is evident from the definition that

$$p_* = g_{\text{rev}}(p_1, \ldots, p_n)$$

for some function $g = g_{\text{rev}}$; however, it is not possible to compute $g$ without making further assumptions on the model. At this level of specificity, the model, while too general to be applied, is almost certainly correct. If we pick a particular probability model $(\Omega, \mathcal{F}, \mathbb{P})$, event $A \in \mathcal{F}$, and sub-$\sigma$-fields $\{\mathcal{F}_i\}$, then the model will almost certainly be wrong but will, at least in principle, determine $g$. Philosophically, we might think of the goal as to choose the model that is least wrong among models in which $g$ may be computed.

### 6.1.3 Gaussian partial information model

In Satopää et al. (2015) the following Gaussian model is introduced. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supports a centered Gaussian process $\{X_A : A \subseteq S\}$ indexed by the Borel subsets of a single set $S$, with $\mathrm{Cov}\,(X_A, X_B) = |A \cap B|$. Here, without loss of generality, $S$ is taken to be the unit interval and $|\cdot|$ refers to Lebesgue measure. The event $A$ is the event that $X_S \geq 0$. The interpretation is that each bit of the white noise $X_{[t,t+dt]}$ adds either positive or negative information about the occurrence of $A$, this being decided in the end by whether the preponderance of information is positive or negative. Each forecaster $\mathcal{F}_i$ is privy to some subset of this information, namely they see all the noise in some subset $B_i \subseteq S$. Formally, $\mathcal{F}_i = \sigma(X_A : A \subseteq B_i)$. Specification of the sets $\{B_i\}$ determines the model and hence $g$.

A number of consequences of this model are worked out in Satopää et al. (2015). The question of how one can efficiently estimate the parameters is taken up in Satopää et al. (2016). In Satopää et al. (2016, Section 5.1), with a slightly enlarged model, these aggregators are tested against data from the Good Judgment Project (Ungar et al., 2012) and compared to the performance of existing aggregators. The parameters of the model necessary for this computation are the unknown covariances between pairs of forecasters. For the purpose of parameter estimation it was important to have not just one forecast but a stream of forecasts for each forecaster.

### 6.1.4 A Bayesian approach to specifying parameters

The present paper considers the problem of applying the Gaussian partial information model in a one-shot forecasting model, i.e., a stream of forecasts is unavailable. The parameters $\{|B_i|, |B_i \cap B_j| : 1 \leq i, j \leq N\}$ cannot consistently be estimated because there is only one data point $p^{(i)}$ for each forecaster $i$. We proceed to model this framework with a Bayesian approach. First, choose a prior $\mu$ on these parameters, chosen to be as uninforma-

tive as possible. Let $\nu$ be the posterior law of the parameters given the forecasts. Then $p_*$ is the mean of $g_\alpha(p^{(1)}, \ldots, p^{(N)})$ when $\alpha$ is an assignment of parameters chosen randomly from the posterior law $\nu$.

Our purpose here is to show that these computations can be carried out for $N = 2$ and one natural choice of prior distribution, and results in an aggregator possessing certain desirable characteristics. Our model is admittedly a toy model, which, on its own, will not beat empirically tuned aggregators. It is hoped, however, that the family of Gaussian-Bayes partial information models includes something close to the correct micro-level model, and that by understanding these models we will understand how to improve on existing non-theoretically based models.

## 6.2 Aggregation function for fixed parameters

We consider a model, as above, where $N = 2$, $|S| = 2$, $|B_1| = |B_2| = 1$ and $|B_1 \cap B_2| = \rho$. The parameter $\rho$, treated in the previous literature as a parameter to be estimated, will later in this paper be taken to be random, uniform on $[0, 1]$. In this section we fix $\rho \in [0, 1]$ and compute the forecast, its marginal distribution and the aggregation function.

### 6.2.1 Computing the forecast and marginals for any parameters

A forecaster observing $X_B$ is ignorant of $X_S - X_B$ which is independent of $X_B$ and has distribution $N(0, |S| - |B|)$ or $\sqrt{|S| - |B|}\chi$ where $\chi$ is a standard normal. Therefore, conditional on $X_B = x$, the forecast is

$$p(x) = \mathbb{P}(X_S - X_B > -x) = \mathbb{P}(\chi < (|S| - |B|)^{-1/2}x) = \Phi\left(\frac{x}{\sqrt{|S| - |B|}}\right).$$

Let $\beta := |B|/(|S| - |B|)$. Because $X_B$ is distributed as $|B|^{1/2}\chi$, we see that the law of $p$ in this model is the law of $\Phi(\beta^{1/2}\chi)$. Because $\chi$ has law $\Phi^{-1}(U)$ for $U$ uniform on $[0, 1]$,

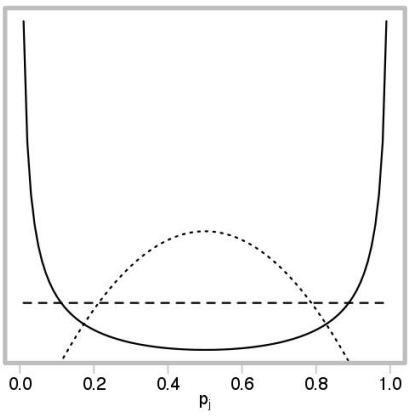



Figure 6.1: The solid line, dashed line, and dotted line are respectively $\beta = 7/3, 1$, and $3/7$

we can summarize this as

$$p(x) \sim \Phi(\beta^{1/2}\Phi^{-1}(U))\,.$$

The density behaves like $(cx(1-x))^{1/\beta}$. When $\beta < 1$ it is unimodal, when $\beta > 1$ it blows up at the endpoints, and when $\beta = 1$ it is exactly uniform; see Figure 6.1.

In this light, the choice of $|B_1| = |B_2| = |S \setminus B_1| = |S \setminus B_2|$ seems natural, as it causes each forecast to be marginally uniform on $[0, 1]$.

### 6.2.2 Computing $g$ under fixed overlap

We now specialize to the Gaussian partial information model $|B_1| = |B_2| = \dfrac{|S|}{2} = 1$ and assume that the parameter $\rho = |B_1 \cap B_2|$ is known. In this section we will compute the aggregating function. In particular:

**Proposition 6.2.1.** *In the Gaussian partial information model with $|B_1| = |B_2| = 1$, $|S| = 2$ and $|B_1 \cap B_2| = \rho$, if the two experts forecast $p^{(1)} = p$ and $p^{(2)} = q$, then the best aggregator $g_\rho(p, q) := \mathbb{P}(A|p^{(1)} = p, p^{(2)} = q)$ is given by*

$$g_\rho(p, q) = \Phi\left(\frac{\Phi^{-1}(p) + \Phi^{-1}(q)}{\sqrt{2\rho(1+\rho)}}\right)\,. \tag{6.3}$$

PROOF: See the Appendix. □

## 6.3 Bayesian model

We enhance the model from 6.2 a Bayesian framework by assuming that the overlap parameter $\rho$ is random, with a prior distribution that is uniform over $[0, 1]$. The posterior distribution is not uniform because the likelihood

$$\lambda_\rho(p, q) := \mathbb{P}(p, q \,|\, \rho)$$

of $(p, q)$ given $\rho$ is nonconstant, whence Bayes' Rule applied with the uniform prior gives a nonconstant posterior. Given $p$ and $q$, posterior probabilities are given by quotients of integrals:

$$\begin{aligned} g(p, q) := \mathbb{P}(A \,|\, p, q) &= \int \mathbb{P}(A \,|\, p, q, \rho) \cdot \mathbb{P}(\rho \,|\, p, q) \\ &= \frac{\int f(p, q, \rho)\lambda_\rho(p, q)\, d\rho}{\int \lambda_\rho(p, q)\, d\rho}. \end{aligned} \tag{6.4}$$

Here we include a factor of $\int \lambda_\rho \, d\rho$ in the denominator so that we may, if we choose, allow $\lambda_\rho$ not to be normalized to have total mass one.

**Theorem 6.3.1.**

$$g(p, q) = \begin{cases} \dfrac{p - (1 - 2q)}{2q} & p > \max\{q, 1 - q\} \\ \dfrac{p}{2(1 - q)} & p < \min\{q, 1 - q\} \\ \dfrac{p - (1 - 2p)}{2p} & q > \max\{p, 1 - p\} \\ \dfrac{q}{2(1 - p)} & q < \min\{p, 1 - p\} \end{cases} \tag{6.5}$$

PROOF: See the Appendix. □

## 6.4 Comparison of Aggregations With Hypothetical Data

For a concrete comparison, suppose two experts forecast respective probabilities $p_1 = 0.6$ and $p_2 = 0.8$. We compare a number of aggregators. The first two were discussed in (6.1) and (6.2), namely the simple average $p^{\text{ave}} := g_{\text{ave}}(p_1, p_2)$ and the probit average $p^{\text{probit}} := g_{\text{probit}}(p_1, p_2)$. As previously discussed, these are constrained to lie between $p_1$ and $p_2$.

We compare the revealed forecast to these two and to two others not constrained to the convex hull. The first of these two is the Gaussian model with fixed overlap parameter $\rho = 1/2$. The second is the log odds summing aggregator. This aggregator, which we have not discussed above, is based on the following probability model. Each forecaster begins with a prior probability estimate of $p = 1/2$ (equivalently $\log(p/(1-p)) = 0$ and sees results of an independent experiment.

By Bayes rule, this experiment affects the posterior probability by some additive increment in the log odds. The result of the two independent experiments is to add both increments to the log odds, resulting in an estimator $p^{\log \text{ odds}}$ which is the most extreme of those we have considered. Just as $p^{\text{ave}}$ and $p^{\text{probit}}$ are demonstrably underconfident, $p^{\log \text{ odds}}$ is overconfident because it assumes that the experts' data are completely disjoint. We then have the following values for the various synthesized forecasts (rounded to the nearest $0.001$).

| | |
|---|---|
| $p^{\text{ave}}$ | $0.700$ |
| $p^{\text{probit}}$ | $0.708$ |
| $p^{1/2}$ | $0.814$ |
| $p^{\text{revealed}}$ | $0.833$ |
| $p^{\log \text{ odds}}$ | $0.857$ |

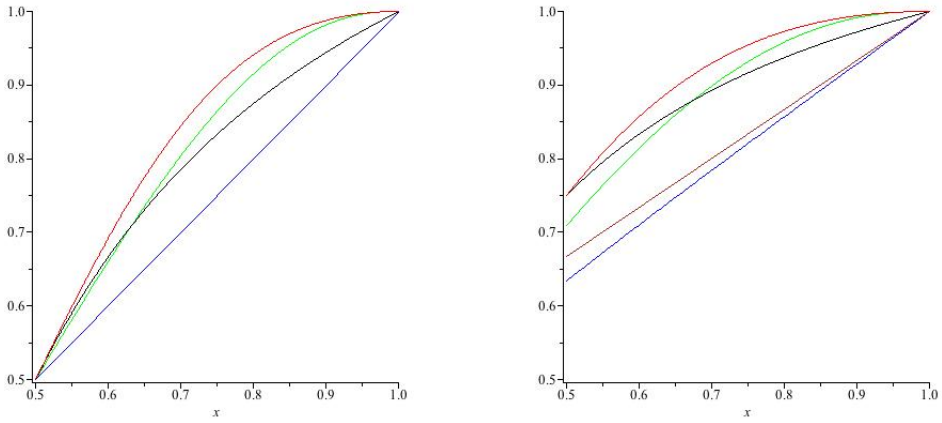



Figure 6.2: Comparisons of aggregators

The first thing we see from this is that the range of values taken by these aggregators is quite broad, extending from $7/10$ at the low end to $6/7$ at the high end[2]. Almost anyone in the business, if given forecasts of $3/5$ and $4/5$ would place their estimate between $7/10$ and $6/7$. The choice of model substantially alters the particular aggregate forecast within the interval of plausible forecasts, and is therefore quite important. We also remark that this choice is not a mathematical one but a practical one. Different forecasting problems may call for different aggregation techniques.

The left graph in Figure 6.2 provides a visual comparison of the above synthesis functions by graphing the diagonal values, that is those where $p = q$. By symmetry, it suffices to graph each of these on the interval $[1/2, 1]$. When $p = q = x$, both the average $p^{\text{ave}}$ and the probit average $p^{\text{probit}}$ are also equal to $x$; these are shown by the blue line. The red curve is $p^{\text{logodds}}$, which is always greatest of the aggregators under consideration. The green and black curves represent $p^{1/2}$ and $p^{\text{revealed}}$ respectively, which are the two partial overlap models. As is evident, these are not strictly ordered. On the right, graphs are shown for $p \in [1/2, 1]$ and $q = (1 + p)/2$. When $p \neq q$, as in the figure on the right, the probit average (shown in brown) is distinct from the average.

One final remark concerns $p^{\text{probit}}$, a popular choice for empirically driven aggregators.

[2]A number of these aggregators give rational values on rational inputs.

While it may seem atheoretical, in fact it arises as the limit as $\rho \to 1$ of the fixed overlap aggregator. To see this, denote the values of $X_S$ as $S$ varies over the algebra of sets generated by $B_1$ and $B_2$ by $U := X_{B_1 \setminus B_2}$, $V := X_{B_2 \setminus B_1}$, $M := X_{B_2 \cap B_1}$ and $W := X_{(B_2 \cup B_1)^c}$; thus $X_{B_1} = U + M, X_{B_2} = V + M$ and $X_S = U + V + M + W$, where $U, V, M, W$ are independent Gaussians with respective variances $1 - \rho, 1 - \rho, \rho, \rho$.

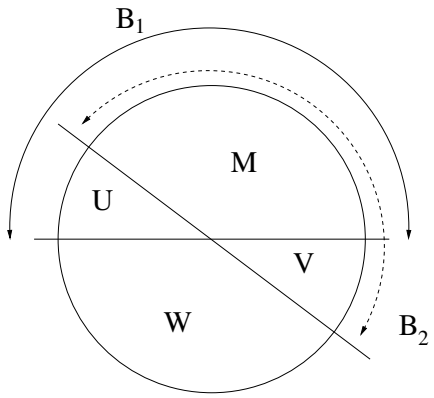



Figure 6.3: Information partition

As $|U| = |V| \to 0$ in Figure 6.3, asymptotically, the likeliest way to achieve $U + M = a$ and $V + M = b$ is to let $M = (a + b)/2$ and $U = -V = (a - b)/2$. These choices become forced in the limit. Applying this with $a = \Phi^{-1}(p)$ and $b = \Phi^{-1}(q)$ shows that the revealed forecast is $\Phi((a + b)/2)$ which is the probit average. In other words, this forecast is practical if we have reason to believe that both forecasters know nearly all information possible (but then somehow mysteriously find highly relevant information in the small part of their information that is not shared).

## 6.5 Comparison of Estimators with 2012 Presidential Election Data

In this section, we apply our proposed Bayesian aggregator from Theorem 6.3.1 to a real data set. The data are collected from two freely online sources: DeSart's 2012 presidential predictions (DeSart, 2015) and Silver's 2012 United States presidential predictions (Silver, 2015). The data are structured as follows: for each state as well District of Columbia, DeSart and Silver give a probability that Obama will win the state's electoral votes. We consider an expert's prediction to be "successful" if the expert-given probability of .5 or greater corresponded to a state in which Obama won (and vice versa for a state in which Obama lost). In 2012, both DeSart and Silver predicted each state's outcome successfully, (the only exception worth noting is that Silver predicted Obama would win Florida with probability .5). To quantify the "success" of each aggregator, we employ ROC (receiver operating curve) analysis. That is, for each classification cutoff value, we calculate the sum of the resulting true positives and the resulting true negatives and divide this number by the total sample size of 51. This is done in Section 6.5.2, in which a ROC analysis is performed for our proposed Bayesian aggregator, the inverse-phi aggregator, and the mean aggregator.

### 6.5.1 Exploratory Data Analysis

When there are two experts leaning over, say, 80% in one direction, as defined, the Bayesian aggregator will be more extreme than either prediction. A scatter plot with the DeSart and Silver predictions along with the Bayesian aggregation model's predictions shown in Figure 6.4 to support this claim. Table 6.2 further supports this claim by displaying a few state indexes (and their corresponding states) that have the highest discrepancy between the Bayesian aggregator and the expert predictors of DeSart and Silver.

In Figure 6.5, we display a scatter plot for the predictions for each state from the

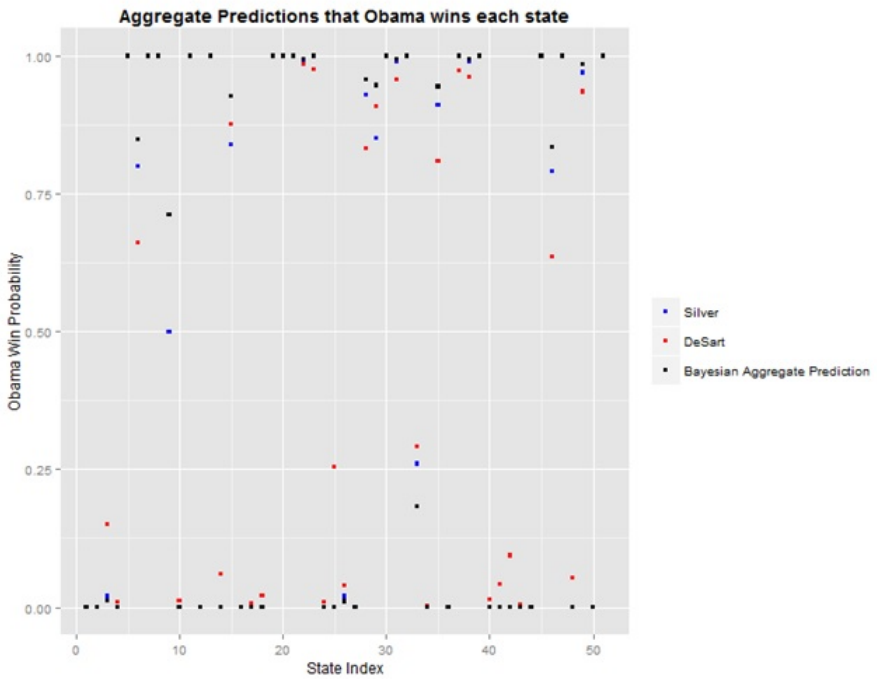



Figure 6.4: DeSart predictions, Silver predictions, and Bayesian aggregation predictions

| State ID | State | Aggregation | DeSart | Silver |
|---|---|---|---|---|
| 3 | Arizona | .012 | .15 | .02 |
| 6 | Colorado | .849 | .662 | .8 |
| 9 | Florida | .712 | .712 | .5 |
| 15 | Iowa | .926 | .876 | .84 |
| 26 | Montana | .011 | .04 | .02 |
| 49 | Wisconsin | .984 | .935 | .97 |

Table 6.1: Bayesian aggregation and expert predictions for individual states

Bayesian model, from the inverse-phi aggregation, and from the raw average aggregation. Notice in Figure 6.5 that the Bayesian aggregate predictions are always more extreme or equally as extreme as the mean and inverse-phi aggregation methods.

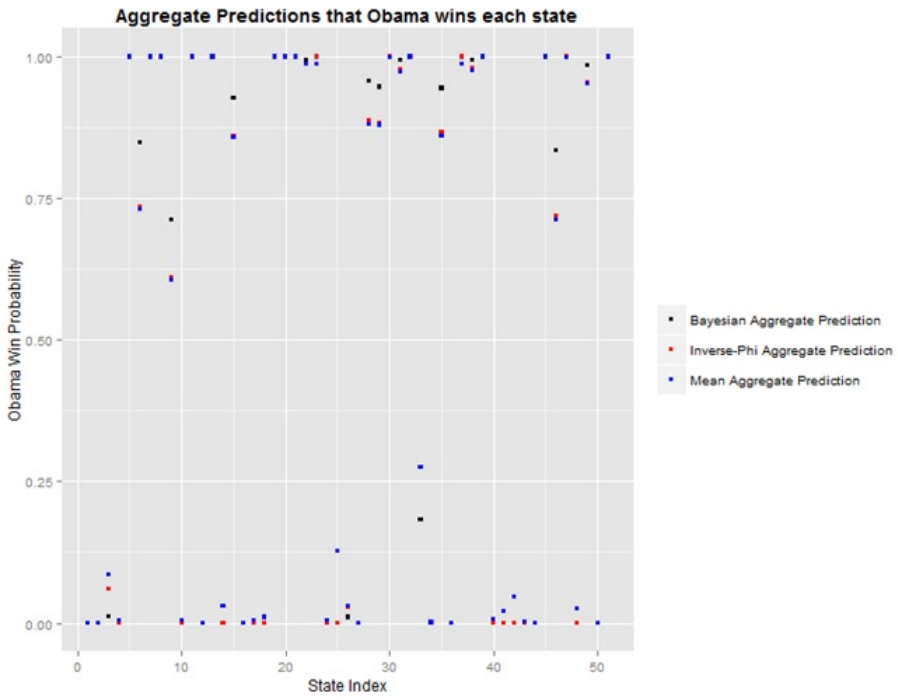



Figure 6.5: Predictions for three aggregation models: Bayesian aggregation, inverse-phi aggregation, and raw average aggregation

We now proceed to calculate the squared-error loss for each aggregation methodology. For each state, we have a prediction, call it $\hat{p}_i$, for each of the three aggregators. We then have the observed data, $Y_i$, corresponding to an Obama victory or loss in that state ($Y_i = 1$ is a win, $Y_i = 0$) is a loss. The sum of the squared error loss $\sum_{i=1}^{51} (\hat{p}_i - Y_i)^2$ is calculated below for the three aggregation methodologies. Note the small squared error loss across for all three aggregators; this is due to the experts' very high accuracy in prediction.

| Bayesian | Inverse | Mean |
|---|---|---|
| .1805 | .4492 | .4875 |

Table 6.2: Squared error loss for aggregation procedures

### 6.5.2 ROC Analysis

In Table 6.3 below we provide a few values from the ROC table for classifying a prediction of Obama victory. In the table, the number True positive (TP) and true negative (TN) predictions for each cutoff value are counted as TP+TN. The maximum value of TP+TN is the number of observations, which in this case is 51.

| Cutoff $p_0$ | Bayesian aggregation (TP+TN) | Inverse-phi aggregation (TP+TN) | Mean-aggregation (TP+TN) |
|---|---|---|---|
| .01 | 27+21=48 | 27+21=48 | 27+15=42 |
| .05 | 27+23=50 | 27+22=49 | 27+21=48 |
| .10 | 27+23=50 | 27+23=50 | 27+22=49 |
| .25 | 27+24=51 | 27+23=50 | 27+23=50 |
| .5 | 27+24=51 | 27+24=51 | 27+24=51 |
| .75 | 26+24=50 | 24+24=48 | 24+24=48 |
| .90 | 24+24=48 | 20+24=44 | 20+24=44 |
| .95 | 21+24=45 | 20+24=44 | 20+24=44 |
| .99 | 19+24=43 | 16+24=40 | 14+24=38 |

Table 6.3: Some rows of ROC table

We see that for cutoff of $p = .5$ (that is, we classify $Y_i = 1$ if $p_i \geq .5$), all three methods have no misclassifications. For the cutoffs $p = 0.75, 0.90, 0.95$ and $0.99$, our Bayesian aggregation method outlined in this paper is superior to both the inverse-phi aggregation and the mean aggregation. We also note that our method has better predictions for every cutoff than the mean aggregation (except for the cutoff $p = 0.5$). It performs equivalently to inverse-phi aggregation for $p = 0.01, 0.1$ and $0.5$ but produces outperforms inverse-phi aggregation for all other cutoff values. These results give us confidence in the performance of our proposed one-shot Bayesian aggregator.

## 6.6 Acknowledgments

This research was supported in part by NSF grant # DMS-1209117 and a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

# 7

# Combining and Extremizing Real-Valued Forecasts*


## Abstract

The weighted average is by far the most popular approach to combining multiple forecasts of some future outcome. This paper shows that both for probability or real-valued forecasts, a non-trivial weighted average of different forecasts is always sub-optimal. More specifically, it is not consistent with any set of information about the future outcome even if the individual forecasts are. Furthermore, weighted averaging does not behave as if it collects information from the forecasters and hence needs to be *extremized*, that is, systematically transformed away from the marginal mean. This paper proposes a linear extremization technique for improving the weighted average of real-valued forecasts. The resulting more extreme version of the weighted average exhibits many properties of optimal aggregation. Both this and the sub-optimality of the weighted average are illustrated with simple examples involving synthetic and real-world data.


*Joint work with Lyle H. Ungar

## 7.1 Introduction

Policy-makers often consult human or/and machine agents for forecasts of some future outcome. For instance, multiple economics experts may provide quarterly predictions of gross domestic product (GDP). Typically it is not possible to determine ex-ante which expert will be the most accurate, and even if this could be done, heeding only the most accurate expert's advice would ignore a potentially large amount of relevant information that is being contributed by the rest of the experts. Therefore a better alternative is to combine the forecasts into a single consensus forecast that represents all the experts' advice. The policy-makers, however, can choose to aggregate the forecasts in many different ways. The final choice of the combination rule is crucial because it often decides how much of the experts' total information is incorporated and hence how well the consensus forecast performs in terms of predictive accuracy.

Possibly because of its simplicity and intuitive appeal, the most popular approach to combining forecasts is the weighted average, sometimes also known as the linear opinion pool. This technique has a long tradition, with many empirical studies attesting to its benefits (see, e.g., Bates and Granger 1969; Clemen 1989; Armstrong 2001). Even though the average forecast does not always outperform the best single forecaster (Hibon and Evgeniou, 2005), it is still considered state-of-the-art (Elliott and Timmermann, 2013) in many fields, including economics (Blix et al., 2001), weather forecasting (Raftery et al., 2005), political science (Graefe et al., 2014b), and many others. In this paper, however, we show that non-trivial weighted averaging is suboptimal, and propose a simple transformation to improve it. A more detailed description of the contributions is given below.

In practice forecasts are typically either real-valued or probabilities of binary events, such as rain or no rain tomorrow. Ranjan and Gneiting (2010) focus on the latter and explain how the quality of a probability forecast (individual or aggregate) is typically measured in terms of *reliability* and *resolution* (sometimes also known as calibration and sharpness, re-

spectively). Reliability describes how closely the conditional event frequencies align with the forecast probabilities. Resolution, on the other hand, measures how far the forecasts are from the naive baseline forecast, that is, the marginal event frequency. A forecast that is reliable and highly resolute is very useful to the policy-maker because it is both accurate and close to the most confident values of zero and one. Therefore a well-established goal in probability forecasting is to maximize resolution subject to reliability (Murphy and Winkler, 1987b; Gneiting et al., 2007).

Strikingly, Ranjan and Gneiting (2010) prove that any non-trivial weighted average of two or more different, reliable probability forecasts is unreliable and lacks resolution. In particular, they explain that such a weighted average is under-confident in a sense that it is overly close to the marginal event frequency. This result is an important contribution to the probability forecasting literature in part because it points out a dramatic shortcoming of methodology that is used widely in practice. However, the authors neither provide a principled way of addressing the shortcoming nor interpret potential causes of the under-confidence.

The first step towards addressing these issues and improving the general practice of aggregation is to understand what is meant by principled aggregation. This topic was discussed by Satopää et al. (2016, 2015) who propose the *partial information framework* as a general platform for modeling and combining forecasts. Under this framework, the outcome and the forecasts share a probability space but without any restrictions on their dependence structure. Any forecast heterogeneity is assumed to stem purely from information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles about the state of the economy may use distinct parts of the information and hence report different predictions of the next quarter's GDP. Even though, to date, this framework has been mainly used for constructing new aggregators, it also offers an ideal environment for analyzing other, already existing, aggregation tech-

niques. No previous work, however, has used it to study weighted averaging of probability or real-valued forecasts.

The first contribution of this paper leaves the type of forecasts unspecified and analyzes the weighted average of any univariate forecasts under the partial information framework. The results are general and encompass both probability and real-valued forecasts. First, the aforementioned result in Ranjan and Gneiting (2010) is generalized to any type of univariate forecasts. This result shows, for instance, that any non-trivial weighted average of reliable predictions about the next quarter's GDP is both unreliable and under-confident. Second, some general properties of optimal aggregation are enumerated. This leads to an original point of view on forecast aggregation, general, yet intuitive, descriptions of well-known properties such as reliability and resolution, and an introduction of a new property, called *variance expansion*, that is associated with aggregators whose variance is never less than the maximum variance among the individual forecasts. Such aggregators are called *expanding* and can be considered to collect information from the individual forecasters. Showing that a non-trivial weighted average is never expanding leads to a mathematically precise yet easy-to-understand explanation of why weighted averages tend to be under-confident. This reasoning suggests that under-confidence is not unique to the class of weighted averages but extends to many other measures of central tendency, such as the median, that also tend to reduce variance.

In probability forecasting the under-confidence of a simple aggregator, such as the average or median, is typically alleviated by a heuristic known as *extremizing*, that is, by systematically transforming the aggregate towards its nearer extreme (at zero or one). For instance, Ranjan and Gneiting (2010) propose a beta transformation that extremizes the weighted average of the probability forecasts; Satopää et al. (2014) use a logistic regression model to extremize the average log-odds of the forecasts; many others, including Shlomi and Wallsten (2010), Baron et al. (2014), and Mellers et al. (2014), have also discussed ex-

tremization of probability forecasts. Intuitively, extremization increases confidence by explicitly moving the aggregate closer to the most confident values of zero and one. Naturally, the same intuition applies to probability forecasts of any categorical outcome. However, if the outcome and forecasts are real-valued, it is not clear anymore what values represent the most confident forecasts. Consequently, it seems that extremization, as described above, lacks direction and cannot be applied. Furthermore, the idea of extremizing may seem counter-intuitive given the large amount of literature attesting to the benefits of shrinkage (James and Stein, 1961). These may be the main reasons why, to the best of our knowledge, no previous literature has discussed extremization of real-valued forecasts.

Therefore it is perhaps somewhat surprising that our second contribution shows that extremizing can improve aggregation also when the individual forecasts are real-valued. First, the notion of extremizing is made precise. This involves introducing a general definition that differs slightly from the above heuristic. In particular, extremizing is redefined as a shift away from the least confident forecast, namely the marginal mean of the outcome, instead of towards the most confident (potentially undefined) values. Second, our definition and theoretical analysis motivate a convex optimization procedure that linearly extremizes the optimally weighted average of real-valued forecasts. The technique is illustrated on simple examples involving both synthetic and real-world data. In each example extremizing leads to improved aggregation with many of the optimal properties enumerated in the beginning of the analysis.

The rest of the paper is structured as follows. Section 7.2 briefly introduces the general partial information framework and discusses some properties of the optimal aggregation within that framework. The class of weighted averages is then analyzed in the light of these properties. Section 7.3 describes the optimization technique for extremizing the weighted average of real-valued forecasts. Section A.4.3 illustrates this technique and our theoretical results over synthetic data. Section 7.5 repeats the analysis over real-world data. The final

section concludes and discusses future research directions.

## 7.2 Forecast and Aggregation Properties

### 7.2.1 Optimal Aggregation

Consider $N$ forecasters and suppose forecaster $j$ predicts $X_j$ for some (integrable) quantity of interest $Y$. The partial information framework assumes that $Y$ and $X_j$, for $j = 1, \ldots, N$, are measurable random variables under some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Akin to Murphy and Winkler (1987b), Ranjan and Gneiting (2010), Jolliffe and Stephenson (2012), and many others, the forecasters are assumed to be *reliable*, that is, conditionally unbiased such that $\mathbb{E}(Y|X_j) = X_j$ for all $j = 1, \ldots, N$. To interpret this assumption, observe that the principal $\sigma$-field $\mathcal{F}$ holds all possible information that can be known about $Y$. Each reliable forecast $X_j$ then generates a sub-$\sigma$-field $\sigma(X_j) := \mathcal{F}_j \subseteq \mathcal{F}$ such that $X_j = \mathbb{E}(Y|\mathcal{F}_j)$. Conversely, suppose that $X_j = \mathbb{E}(Y|\mathcal{F}_j)$ for some $\mathcal{F}_j \subseteq \mathcal{F}$, then

$$\mathbb{E}(Y|X_j) = \mathbb{E}[\mathbb{E}(Y|X_j, \mathcal{F}_j)|X_j] = \mathbb{E}[\mathbb{E}(Y|\mathcal{F}_j)|X_j] = \mathbb{E}(X_j|X_j) = X_j.$$

Therefore a forecast is reliable if and only if it represents the optimal use of some information set, that is, it is consistent with some partial information $\mathcal{F}_j \subseteq \mathcal{F}$. Given that at this level of specificity the framework is highly general and hence likely to be a good approximation of real-world prediction polling, it offers an ideal platform for analyzing different aggregators.

In this paper an aggregator is defined to be any forecast that is measurable with respect to $\mathcal{F}'' := \sigma(X_1, \ldots, X_N)$, namely the $\sigma$-field generated by the individual forecasts. For the sake of notational clarity, aggregators are denoted with different versions of the script symbol $\mathcal{X}$. If $\mathbb{E}(Y^2) < \infty$, the conditional expectation $\mathcal{X}'' := \mathbb{E}(Y|\mathcal{F}'')$ minimizes the

expected quadratic loss among all aggregators (see, e.g., Durrett 2010). This forecast is called the *revealed aggregator* because it optimally utilizes all the information that the forecasters' reveal through their forecasts. Even though $\mathcal{X}''$ is typically too abstract to be applied in practice, it provides an optimal baseline for aggregation efficiency. Therefore studying its properties gives guidance for improving aggregators currently used in practice. Some of these properties are summarized in the following theorem. The proof is deferred to the Appendix.

**Theorem 7.2.1.** *Suppose that $X_j = \mathbb{E}(Y|X_j)$ for all $j = 1, \ldots, N$ and denote the revealed aggregator with $\mathcal{X}'' = \mathbb{E}(Y|\mathcal{F}'')$, where $\mathcal{F}'' = \sigma(X_1, \ldots, X_N)$. Let $\delta_{max} := \max_j\{\mathrm{Var}\,(X_j)\}$ be the maximal variance among the individual forecast. Then the following holds.*

*i)* ***Marginal Consistency.*** *$\mathcal{X}''$ is marginally consistent: $\mathbb{E}(\mathcal{X}'') = \mathbb{E}(Y) := \mu_0$.*

*ii)* ***Reliability.*** *$\mathcal{X}''$ is reliable: $\mathbb{E}(Y|\mathcal{X}'') = \mathcal{X}''$.*

*iii)* ***Variance Expansion.*** *$\mathcal{X}''$ is expanding: $\delta_{max} \leq \mathrm{Var}\,(\mathcal{X}'')$. In words, the variance of $\mathcal{X}''$ is always at least as large as that of the most variable forecast.*

Marginal consistency states that the forecast and the outcome agree in expectation. If $X_j$ is reliable, then $\mathbb{E}(X_j) = \mathbb{E}[\mathbb{E}(Y|X_j)] = \mathbb{E}(Y) = \mu_0$. Consequently, all reliable forecasts (individual or aggregate) are marginally consistent. The converse, however, is not true. For instance, Theorem 7.2.2 (see Section 7.2.2) shows that any non-trivial weighted average is marginally consistent but unreliable. This is an important observation because it provides a technique for proving lack of reliability via marginal inconsistency – a task that is generally much easier than disproving reliability directly.

Given that each reliable forecast can be associated with a sub-$\sigma$-field and that conditional expectation is a contraction in $L^2$ (Durrett, 2010, Theorem 5.1.4.), the variance of

any reliable forecast (individual or aggregate) is always upper-bounded by $\mathrm{Var}\,(Y)$. Theorem 7.2.1 further shows that the corresponding lower bound for $\mathrm{Var}\,(\mathcal{X}'')$ is the maximum variance among the forecasters. To interpret this lower bound, consider an increasing sequence of $\sigma$-fields $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_R \subseteq \mathcal{F}$ and the corresponding forecasts $X_r = \mathbb{E}(Y|\mathcal{F}_r)$ for $r = 0, 1, \ldots, R$. According to Satopää et al. (2016, Proposition 2.1), the variances of these forecasts respect the same order as their information sets: $\mathrm{Var}\,(X_0) \leq \mathrm{Var}\,(X_1) \leq \cdots \leq \mathrm{Var}\,(X_R) \leq \mathrm{Var}\,(Y)$. This suggests that the amount of information used in a reliable forecast is reflected in its variance. Naturally, if an aggregator collects information from a group of forecasters, it should use at least as much information as the most informed individual forecaster; that is, its variance should exceed that of the individual forecasters'. Therefore any aggregator that *expands* variance and satisfies this condition is considered a collector of information.

Recall that in probability forecasting a well-established goal is to maximize resolution subject to reliability. This goal can be easily interpreted intuitively with the help of partial information. First, conditioning on reliability requires the forecast to be consistent with some set of information about $Y$. Maximizing the resolution of this forecast takes it as far from $\mu_0$ as possible. This is equivalent to increasing the variance of the forecast as close to the theoretical upper bound $\mathrm{Var}\,(Y)$ as possible. Therefore the goal is equivalent to maximizing the amount of information that the forecast is consistent with. Intuitively, this is very reasonable and should be considered as the general goal in forecasting.

### 7.2.2 Weighted Averaging

The rest of the paper analyzes the most commonly used aggregator, namely the weighted average. The following theorem shows that a non-trivial weighted average is neither expanding nor reliable and therefore can be considered suboptimal. The proof is again deferred to the Appendix. A similar result does not hold for all linear combinations of the

individual forecasts. For instance, Section A.4.3 describes a model under which the optimal aggregator $\mathcal{X}''$ is always a linear combination of the individual $X_j$'s.

**Theorem 7.2.2.** *Suppose that $X_j = \mathbb{E}(Y|X_j)$ for $j = 1, \ldots, N$. Denote the weighted average with $\mathcal{X}_w := \sum_{j=1}^{N} w_j X_j$, where $w_j \geq 0$, for all $j = 1, \ldots, N$, and $\sum_{j=1}^{N} w_j = 1$. Let $m = \arg\max_j \{\mathrm{Var}\,(X_j)\}$ identify the forecast with the maximal variance $\delta_{max} = Var(X_m)$. Then the following holds.*

*i) $\mathcal{X}_w$ is marginally consistent.*

*ii) $\mathcal{X}_w$ is not reliable, that is, $\mathbb{P}\left[\mathbb{E}(Y|\mathcal{X}_w) \neq \mathcal{X}_w\right] > 0$ if there exists a forecast pair $i \neq j$ such that $\mathbb{P}(X_i \neq X_j) > 0$ and $w_i, w_j > 0$. In words, $\mathcal{X}_w$ is necessarily unreliable if it assigns positive weight to at least two different forecasts.*

*iii) Under the conditions of item ii), $\mathcal{X}_w$ lacks resolution. More specifically, if $\mathcal{X}'_w := \mathbb{E}(Y|\mathcal{X}_w)$ is the reliable version of $\mathcal{X}_w$, then $\mathbb{E}(\mathcal{X}_w) = \mathbb{E}(\mathcal{X}'_w) = \mu_0$ but $\mathrm{Var}\,(\mathcal{X}_w) < \mathrm{Var}\,(\mathcal{X}'_w)$. In other words, $\mathcal{X}_w$ is under-confident in a sense that it is closer to the marginal mean $\mu_0$ than its reliable version $\mathcal{X}'_w$.*

*iv) $\mathcal{X}_w$ is not expanding. In particular, $\mathrm{Var}\,(\mathcal{X}_w) \leq \delta_{max}$, which shows that $\mathcal{X}_w$ is under-confident in a sense that it is as close or closer to the marginal mean $\mu_0$ than the revealed aggregator $\mathcal{X}''$. Furthermore, $\mathrm{Var}\,(\mathcal{X}_w) = \mathrm{Var}\,(\mathcal{X}'')$ if and only if both $\mathcal{X}_w = \mathcal{X}'' = X_m$; that is, $X_m$ provides all the information necessary for $\mathcal{X}''$, and $\mathcal{X}_w$ assigns all weight to $X_m$ (or to a group of forecasts all equal to $X_m$).*

This theorem discusses under-confidence under two different baselines. Item iii) is a generalization of Ranjan and Gneiting (2010, Theorem 2.1.). Intuitively, it states that if $\mathcal{X}_w$ is trained to use its information accurately, the resulting aggregator is more confident. Therefore under-confidence is defined relative to the reliable version of $\mathcal{X}_w$. Under this kind of comparison, however, a reliable aggregator is never under-confident. For instance,

an aggregator that ignores the individual forecasts and always returns the marginal mean $\mu_0$ is reliable and hence would not be considered under-confident. Intuitively, however, it is clear that no aggregate forecast is more under-confident than the marginal mean $\mu_0$. To address this drawback, item iv) defines under-confidence relative to the revealed aggregator instead. Such a comparison estimates whether the weighted average is as confident as it should be given the information it received through the forecasts. Item iv) shows that this happens only if all the weight is assigned to a forecaster whose information set contains every other forecasters' information. However, even if $\mathcal{X}_w$ could pick out the most informed forecaster ex-ante, the chances of a single forecaster knowing everything that the rest of the forecasters know is extremely small in practice. In essentially all other cases, $\mathcal{X}_w$ is under-confident, unreliable, and hence not consistent with some set of information about $Y$.

Unfortunately, this shortcoming spans across all measures of central tendency. These aggregators reduce variance and hence are separated from the revealed aggregator by the maximum variance among the individual forecasts. For instance, Papadatos (1995) discuss the maximum variance of different order statistics and show that the variance of the median is upper bounded by the global variance of the individual forecasts. Given that such aggregators are not expanding, they cannot be considered to collect information. To illustrate, consider a group of forecasters, each independently making a probability forecast of $0.9$ for the occurrence of some future event. If these forecasters are using different evidence, then clearly the combined evidence should give an aggregate forecast somewhat greater than $0.9$. In this simple scenario, however, measures of central tendency will always aggregate to $0.9$. Therefore they fail to account for the information heterogeneity among the forecasters. Instead, they reduce "measurement error," which is philosophically very different to the idea of information aggregation discussed in this paper.

Theorem 7.2.2, however, is not only negative in nature; it is also constructive in several different ways. First, it motivates a general and precise definition of extremizing:

**Definition 7.2.3. Extremization.** Consider two reliable forecasts $X_i$ and $X_j$. Denote their common marginal mean with $\mathbb{E}(X_i) = \mathbb{E}(X_j) = \mu_0$. The forecast $X_j$ *extremizes* $X_i$ if and only if either $X_j \leq X_i \leq \mu_0$ or $\mu_0 \leq X_i \leq X_j$ always holds.

It is interesting to contrast this definition with the popular extremization heuristic in the context of probability forecasting. Definition 7.2.3 suggests that simply moving, say, the average probability forecast closer to zero or one improves the aggregate if and only if the marginal probability of success is $0.5$. In other cases naively following the heuristic may end up degrading the aggregate. For instance, consider a geographical region where rain is known to occur on $20\%$ of the days. If the average probability forecast of rain tomorrow is $0.30$, instead of following the heuristic and shifting this aggregate towards zero and hence closer to the marginal mean of $0.20$, the aggregate should be actually shifted in the opposite direction, namely closer to one. Second, Theorem 7.2.2 suggests that extremization, as defined formally above, is likely to improve the weighted average of any type of univariate forecasts. This justifies the construction of a broader class of extremizing techniques. In particular, the second part of item iv) states that extremizing is likely to improve the weighted average when the single most informed forecaster knows a lot less than all the forecasters know as a group. To illustrate this, the next section introduces a simple optimization procedure that extremizes the weighted average of real-valued forecasts.

## 7.3 Extremizing Real-Valued Forecasts

Estimating the weights and the amount of extremization requires the forecasters to address more than one related problems. For instance, they may participate in separate yet similar prediction problems or give repeated forecasts on a single recurring event. Across such problems the weights and the resulting under-confidence are likely to remain stable, allowing the aggregator parameters to be estimated based on multiple predictions per forecaster.

Therefore, from now on, suppose that the forecasters address $K \geq 2$ problems. Denote the outcome of the $k$th problem with $Y_k \in \mathbb{R}$ and let $X_{jk} \in \mathbb{R}$ represent the $j$th forecaster's prediction for this outcome.

Extremization requires at least two parameters: the marginal mean, which acts as the pivot point and decides the direction of extremizing, and the amount of extremization itself. Extremization, of course, could be performed in many different ways. However, if $\mathcal{X}_k^*$ denotes the extremized version of the weighted average for the $k$th problem, then probably the simplest and most natural starting point is the following:

$$\mathcal{X}_k^* = \alpha \left( \mathbf{w}' \mathbf{X}_k - \mu_0 \right) + \mu_0,$$

where $\mathbf{X}_k = (X_{1k}, \ldots, X_{Nk})'$ collects the forecasts for the $k$th outcome, $\mathbf{w} = (w_1, \ldots, w_N)'$ is the weight vector, and $\alpha \in (1, \infty)$ (or $\alpha \in [0, 1)$) leads to extremization (or contraction towards $\mu_0$, respectively). If $\alpha = 1$, then $\mathcal{X}^*$ is equal to the weighted average $\mathcal{X}_w$. This linear form is particularly convenient because it leads to efficient parameter estimation and also maintains marginal consistency of $\mathcal{X}_w$; that is, $\mathbb{E}(\mathcal{X}^*) = \mu_0$ for all values of $\alpha$. However, $\mathrm{Var}\left(\mathcal{X}^*\right)$ increases in $\alpha$ such that $\mathrm{Var}\left(\mathcal{X}^*\right) = \alpha^2 \mathrm{Var}\left(\mathcal{X}_w\right) > \mathrm{Var}\left(\mathcal{X}_w\right)$ for all $\alpha > 1$. Therefore, for a large enough $\alpha$, $\mathcal{X}^*$ is both marginally consistent and expanding. These properties hold even if the weighted average is replaced by some other marginally consistent aggregator. However, given that the main purpose of this procedure is to illustrate Theorem 7.2.2, this paper only considers the weighted average.

Recall that the forecasts are assumed calibrated and hence marginally consistent with the outcomes. Therefore an unbiased estimator of the prior mean $\mu_0$ is given by the average of the forecasts $\frac{1}{NK} \sum_{k=1}^{K} \sum_{j=1}^{N} X_{jk}$ or, alternatively, by the average of the outcomes $\frac{1}{K} \sum_{k=1}^{K} Y_k$. Estimating $\mu_0$ in this manner, however, leads to a two-step estimation procedure. A more direct approach is to estimate all the parameters, namely $\alpha$, $\mu_0$, and $\mathbf{w}$, jointly over some criterion. If $Y_k$ has an explicit likelihood in terms of $\mathcal{X}^*$, then the pa-

rameters can be estimated by maximizing this likelihood. Assuming an explicit parametric form, however, can be avoided by recalling from Section 7.2.2 that the revealed aggregator $\mathcal{X}''$ utilizes the forecasters' information optimally and minimizes the expected quadratic loss among all functions measurable with respect to $\mathcal{F}''$. Ideally, $\mathcal{X}^*$ would behave similarly to $\mathcal{X}''$. Therefore it makes sense to estimate its parameters by minimizing the average quadratic loss over some training set. Section A.4.3 shows that this is likely to improve both the resolution and reliability of the weighted average.

These considerations lead to the following estimation problem:

$$\begin{aligned} &\text{minimize } \sum_{k=1}^{K} \left[\alpha \left(\mathbf{w}'\mathbf{X}_k - \mu_0\right) + \mu_0 - Y_k\right]^2 \\ &\text{subject to } w_j \geq 0 \text{ for } j = 1, \ldots, N, \\ &\qquad \sum_{j=1}^{N} w_j = 1, \text{ and} \\ &\qquad \alpha \geq 0. \end{aligned} \tag{7.1}$$

To express this problem in a form that is more amenable to estimation, denote an $N \times N$ identity matrix with $\mathbf{I}_N$, a vector of $K$ ones with $\mathbf{1}_K$, and a vector of $N$ zeros with $\mathbf{0}_N$. If $\mathbf{Y} = (Y_1, \ldots, Y_K)'$, $\mathbf{X} = (\mathbf{1}_K, (\mathbf{X}_1, \ldots, \mathbf{X}_K)')$, and $\mathbf{A} = (\mathbf{0}_N, \mathbf{I}_N)$, then problem (7.1) is equivalent to

$$\begin{aligned} &\text{minimize } \frac{1}{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} \\ &\text{subject to } -\mathbf{A}\boldsymbol{\beta} \leq \mathbf{0}_N, \end{aligned} \tag{7.2}$$

where the inequality is interpreted element-wise and $\boldsymbol{\beta}$ is a vector of $N + 1$ optimization parameters. Given that $\mathbf{X}'\mathbf{X}$ is always positive semidefinite, problem (7.2) is a convex quadratic program that can be solved efficiently with standard optimization techniques. If $\boldsymbol{\beta}^* = (\beta_0^*, \ldots, \beta_N^*)'$ represents the solution to (7.2), the optimal values of the original

parameters can be recovered by

$$\alpha^* = \sum_{j=1}^{N} \beta_j^*,$$
$$w_j^* = \beta_j^*/\alpha^* \text{ for } j = 1, \ldots, N, \text{ and}$$
$$\mu_0^* = -\beta_0^*/(1-\alpha^*).$$

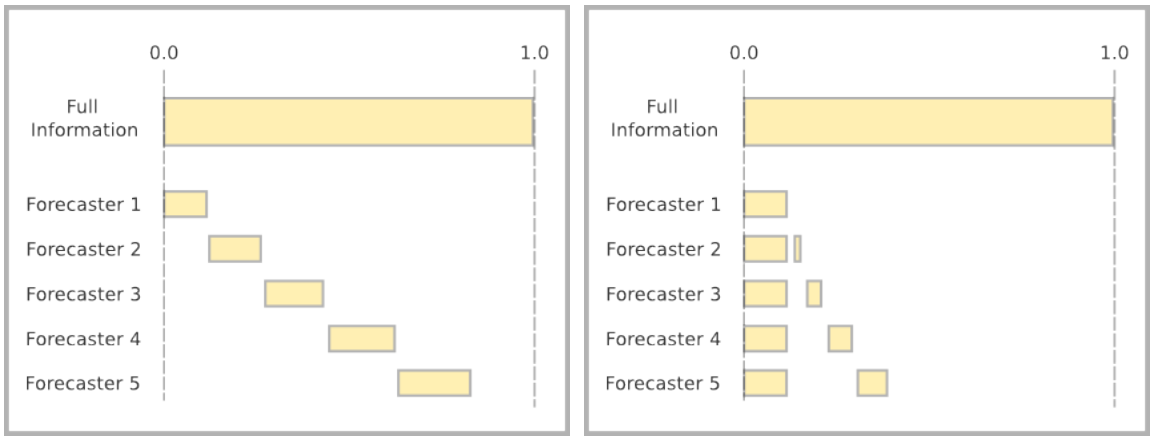
The next two sections apply and evaluate this method both on simulated and real-world data.

## 7.4 Simulation Study

This section illustrates Theorem 7.2.2 on data generated from the Gaussian partial information model introduced in Satopää et al. (2016, 2015) as a close yet practical specification of the general partial information framework. The simplest version of this model occurs when the outcome $Y$ and the forecasts $X_j$ are real-valued with mean zero. The observables for the $k$th problem are then generated jointly from the following multivariate Gaussian distribution:

$$\begin{pmatrix} Y_k \\ X_{1k} \\ \vdots \\ X_{Nk} \end{pmatrix} \sim \mathcal{N}_{N+1}\left(\mathbf{0}, \begin{pmatrix} 1 & \text{diag}(\boldsymbol{\Sigma})' \\ \text{diag}(\boldsymbol{\Sigma}) & \boldsymbol{\Sigma} \end{pmatrix} := \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \ldots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \ldots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \ldots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \ldots & \delta_N \end{array}\right)\right), \tag{7.3}$$

where the covariance matrix describes the *information structure* among the forecasters. In particular, the maximum amount of information is $1.0$. The diagonal entry $\delta_j \in [0, 1]$




(a) No Information Overlap

(b) High Information Overlap

Figure 7.1: Information Distribution Among $N = 5$ Forecasters. The top bar next to *Full Information* represents all possible information that can be known about $Y$. The bar leveled horizontally with *Forecaster* $j$ represents the information used by that forecaster.

represents the amount of information used by forecaster $j$ such that if $\delta_j = 1$ (or $\delta_j = 0$), the forecaster always reports the correct answer $Y_k$ (or the marginal mean $\mu_0 = 0$, respectively). The off-diagonal $\rho_{i,j}$, on the other hand, can be regarded as the amount of information overlap between forecasters $i$ and $j$. Using the well-known properties of a conditional multivariate Gaussian distribution, Satopää et al. (2016, 2015) show that under this model the forecasts are reliable and that the revealed aggregator for the $k$th problem is $\mathcal{X}''_k = \mathbb{E}(Y_k|\mathbf{X}_k) = \text{diag}(\boldsymbol{\Sigma})'\boldsymbol{\Sigma}^{-1}\mathbf{X}_k$.

The distribution (7.3) is particularly useful because it provides a realistic model for testing aggregation under different information structures. This section considers $N = 5$ forecasters under two different structures:

> *No Information Overlap.* Fix $\delta_j = 0.1 + 0.02j$ for $j = 1, \ldots, 5$ and let $\rho_{i,j} = 0$ for all $i, j$. Therefore the forecasters have independent information sources. This information structure is illustrated in Figure 7.1a. Summing up the individual variances shows that as a group the forecasters know $80\%$ of the total information. The

revealed aggregator reduces to $\mathcal{X}_k'' = \sum_{j=1}^{5} X_{jk}$, has variance $0.80$, and therefore efficiently uses all the forecasters' information.

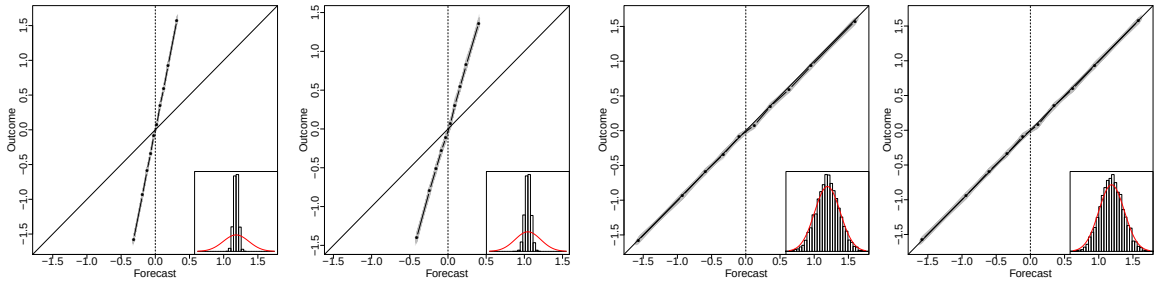
*High Information Overlap.* Fix $\delta_j = 0.1 + 0.02j$ for $j = 1, \ldots, 5$ and let $\rho_{i,j} = 0.12$ for all $i, j$. Therefore the forecasters have significant information overlap and as a group know only $32\%$ of the total information. This information structure is illustrated in Figure 7.1b. The revealed aggregator reduces to $\mathcal{X}_k'' = \left(\sum_{j=2}^{5} X_{jk}\right) - 3X_{1k}$, has variance $0.32$, and therefore efficiently uses all the forecasters' information.

The competing aggregators are the equally weighted average $\bar{\mathcal{X}}$, the optimally weighted average $\mathcal{X}_w$, the extremized version of the optimally weighted average $\mathcal{X}^*$, and the revealed aggregator $\mathcal{X}''$. The parameters in $\mathcal{X}^*$ and $\mathcal{X}_w$ are first estimated by minimizing the average quadratic loss over a training set of $10,000$ draws from (7.3). After this, all the competing aggregators are evaluated on an independent test set of another $10,000$ draws from (7.3). Therefore all the following results, apart from the parameter estimates, represent out-of-sample performance.

In probability forecasting the quality of the predictions is typically assessed using a reliability diagram. The idea is to first sort the outcome-forecast pairs into some number of bins based on the forecasts and then plot the average forecast against the average outcome within each bin. Figures 7.2 and 7.3 generalize this to continuous outcomes by replacing the conditional empirical event frequency with the conditional average outcome. The bins are chosen so that they all contain the same number of forecast-outcome pairs. The vertical dashed line represents the marginal mean $\mu_0 = 0$. The plots have been scaled such that the identity function shows as the diagonal. Any deviation from this diagonal suggests lack of reliability. The grey area represents the reliability diagrams of a $1,000$ bootstrap samples of the forecast-outcome pairs. Therefore it serves as a visual guide for assessing uncertainty. The inset histograms help to assess resolution by comparing the empirical distribution of the forecasts against the prior distribution of $Y$, namely the standard Gaussian distribution

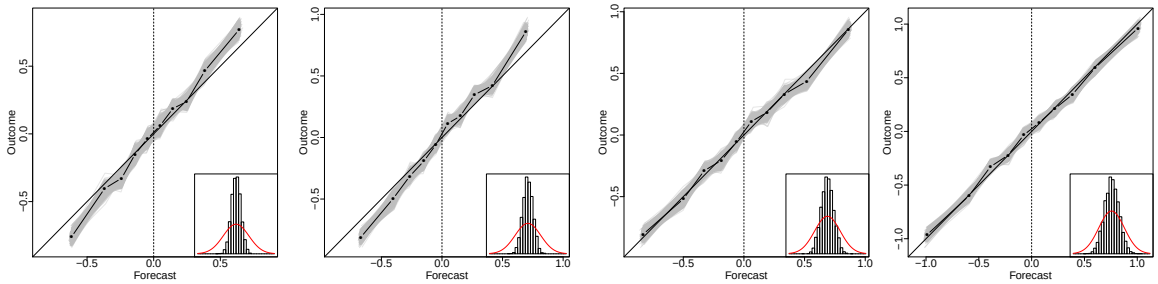
Table 7.1: Synthetic Data. Estimated parameter values.

| Scenario | Forecast | $\mu_0$ | $\alpha$ | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ |
|---|---|---|---|---|---|---|---|---|
| No Overlap | $\mathcal{X}_w$ | | | 0.0000 | 0.1080 | 0.2293 | 0.3025 | 0.3601 |
| | $\mathcal{X}^*$ | 0.0004 | 5.0137 | 0.1964 | 0.2023 | 0.2008 | 0.2006 | 0.2000 |
| High Overlap | $\mathcal{X}_w$ | | | 0.0000 | 0.0000 | 0.0440 | 0.4262 | 0.5298 |
| | $\mathcal{X}^*$ | -0.0077 | 1.3048 | 0.0000 | 0.0000 | 0.1456 | 0.3959 | 0.4585 |




(a) $\bar{\mathcal{X}}$ (b) $\mathcal{X}_w$ (c) $\mathcal{X}^*$ (d) $\mathcal{X}''$

Figure 7.2: Synthetic Data. Out-of-sample reliability under no information overlap.




(a) $\bar{\mathcal{X}}$ (b) $\mathcal{X}_w$ (c) $\mathcal{X}^*$ (d) $\mathcal{X}''$

Figure 7.3: Synthetic Data. Out-of-sample reliability under high information overlap.

represented by the red curve. In particular, if the forecast is reliable, then the closer its empirical distribution is to the standard Gaussian, the more information is being used in the forecast.

Figures 7.2d and 7.3d present the reliability diagrams for $\mathcal{X}''$ under no and high information overlap, respectively. Comparing these plots to the corresponding reliability diagrams of $\bar{\mathcal{X}}$ and $\mathcal{X}_w$ in the same figures, reveals that $\bar{\mathcal{X}}$ and $\mathcal{X}_w$ are not only unreliable but

also have smaller variance than $\mathcal{X}''$. Furthermore, the manner in which the plotted points deviate from the diagonal suggests that $\bar{\mathcal{X}}$ and $\mathcal{X}_w$ are under-confident in both information scenarios. The level of under-confidence is particularly startling in Figures 7.2a and 7.2b but decreases as information overlap is introduced in Figures 7.3a and 7.3b. Given that averaging-like techniques do not behave like information aggregators, that is, they are not expanding, it is not surprising to see them perform better under high information overlap when aggregating information is less important for good performance. Table 7.1 shows the parameter estimates for $\mathcal{X}_w$ and $\mathcal{X}^*$. The weights in $\mathcal{X}_w$ increase in the forecaster's amount of information and differ noticeably from the equal weights employed by $\bar{\mathcal{X}}$. More importantly, however, in both information scenarios $\alpha > 1$. This reflects the need to correct the under-confidence of $\mathcal{X}_w$. The resulting $\mathcal{X}^*$ is more reliable and confident as can be seen in Figures 7.2c and 7.3c. Furthermore, it behaves very similarly to the optimal aggregator $\mathcal{X}''$ under both information structures.

In addition to performing visual assessment, the aggregators can be compared based on their out-of-sample average quadratic loss. To make this specific, let $\mathbf{Y} = (Y_1, \ldots, Y_K)$ collect all the outcomes of the testing problems and $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_K)$ be a vector of some aggregate forecasts for the same problems. Then, the average quadratic loss for this aggregator is

$$L\left(\mathbf{Y}, \boldsymbol{\mathcal{X}}\right) = \frac{1}{K} \sum_{k=1}^{K} \left(Y_k - \mathcal{X}_k\right)^2 .$$

If the forecasts are probability estimates of binary outcomes, the above loss is known to have a decomposition that permits a closer analysis of reliability and resolution (Brier, 1950; Murphy, 1973). The decomposition, however, is not limited to probability forecasts. To see this, suppose that the real-valued aggregate $\mathcal{X}_k \in \{f_1, \ldots, f_I\}$ for some finite number $I$. Let $K_i$ be the number of times $f_i$ occurs, $\bar{Y}_i$ be the empirical average of

Table 7.2: Synthetic Data. The average quadratic loss, $L(\mathbf{Y}, \boldsymbol{\mathcal{X}})$ with its three additive components: reliability (REL), resolution (RES), and uncertainty (UNC). The final column, $s^2$ gives the estimated variance of the forecast.

| Scenario | Forecast | $L(\mathbf{Y}, \boldsymbol{\mathcal{X}})$ | REL | RES | UNC | $s^2$ |
|---|---|---|---|---|---|---|
| No Overlap | Best Individual | 0.8024 | 0.0050 | 0.2108 | 1.0081 | 0.200 |
| | Median | 0.7322 | 0.2928 | 0.5688 | 1.0081 | 0.046 |
| | $\bar{\mathcal{X}}$ | 0.7185 | 0.5140 | 0.8036 | 1.0081 | 0.032 |
| | $\mathcal{X}_w$ | 0.7016 | 0.2913 | 0.5979 | 1.0081 | 0.055 |
| | $\mathcal{X}^*$ | 0.1971 | 0.0022 | 0.8132 | 1.0081 | 0.799 |
| | $\mathcal{X}''$ | 0.1969 | 0.0021 | 0.8132 | 1.0081 | 0.807 |
| High Overlap | Best Individual | 0.8141 | 0.0061 | 0.2195 | 1.0275 | 0.199 |
| | Median | 0.8492 | 0.0087 | 0.1870 | 1.0275 | 0.125 |
| | $\bar{\mathcal{X}}$ | 0.8254 | 0.0137 | 0.2157 | 1.0275 | 0.128 |
| | $\mathcal{X}_w$ | 0.7889 | 0.0166 | 0.2552 | 1.0275 | 0.150 |
| | $\mathcal{X}^*$ | 0.7758 | 0.0056 | 0.2573 | 1.0275 | 0.228 |
| | $\mathcal{X}''$ | 0.6837 | 0.0057 | 0.3496 | 1.0275 | 0.318 |

$\{Y_k : \mathcal{X}_k = f_i\}$, and $\bar{Y} = \frac{1}{K}\sum_{k=1}^{K} Y_k$. Then,

$$L(\mathbf{Y}, \boldsymbol{\mathcal{X}}) = \underbrace{\frac{1}{K}\sum_{i=1}^{I} K_i(f_i - \bar{Y}_i)^2}_{\text{REL}} - \underbrace{\frac{1}{K}\sum_{i=1}^{I} K_i(\bar{Y}_i - \bar{Y})^2}_{\text{RES}} + \underbrace{\frac{1}{K}\sum_{k=1}^{K}(Y_k - \bar{Y})^2}_{\text{UNC}}. \tag{7.4}$$

See the Appendix for the derivation of this decomposition. The three components of the decomposition are highly interpretable. In particular, low REL suggests high reliability. If the aggregate is reliable, then RES is approximately equal to the sample variance of the aggregate and is increasing in resolution. The final term, UNC does not depend on the forecasts. This is the sample variance of $Y$ and therefore gives an approximate upper bound on the variance of any reliable forecast. As has been mentioned before, the goal is to maximize resolution subject to reliability. This decomposition shows how the quadratic loss addresses reliability and resolution simultaneously and therefore provides a convenient loss function for learning aggregation parameters.

Table 7.2 presents the quadratic loss, its additive components, and the estimated variance $s^2$ for each of the different forecasts under both information scenarios. In addition to the aforementioned $\bar{\mathcal{X}}$, $\mathcal{X}_w$, $\mathcal{X}^*$, and $\mathcal{X}''$, the table also presents scores for the median forecast and the individual forecaster with the lowest quadratic loss. Even though the best individual is reliable by construction, it is highly unresolute and hence gains an overall poor quadratic loss. Under no information overlap, however, this individual is better than both the median and $\bar{\mathcal{X}}$ because these aggregators assign too much importance to the individual forecasters with very little information. As predicted by Theorem 7.2.2, the median and the averaging aggregators $\bar{\mathcal{X}}$ and $\mathcal{X}_w$ are neither reliable nor expanding. The remaining two aggregators, namely $\mathcal{X}^*$ and $\mathcal{X}''$, on the other hand, are reliable and expanding. Table 7.1 shows that $\mathcal{X}^*$ is in fact almost equivalent to $\mathcal{X}''$ under no information overlap. Under high information overlap, however, $\mathcal{X}''$ gains slight advantage over $\mathcal{X}^*$. In this case $\mathcal{X}^*$ cannot take the same form as $\mathcal{X}''$. Consequently, it has an estimated variance of $0.228$ which is well below the amount of information known to the group, namely $0.320$. It fails to use information optimally because it cannot subtract off the shared information $X_1$ and hence avoid double-counting of information. However, despite it using information less efficiently, it is as reliable as $\mathcal{X}''$.

Of course, under the Gaussian model, $\mathcal{X}^*$ may seem redundant because the optimal $\mathcal{X}''$ can be computed directly. In practice, however, $\mathbf{\Sigma}$ is not known and must be estimated under a non-trivial semidefinite constraint (see Satopää et al. 2016 for more details). Given that this involves a total of $\binom{N}{2} + N$ parameters, the estimation task is challenging even for moderately large $N$, say, greater than $100$. Furthermore, accurately estimating such a large number of parameters requires the forecasters to attend a large number of prediction problems. Applying $\mathcal{X}^*$ instead is significantly easier because it involves only $N + 1$ parameters that can be estimated via a standard quadratic program (7.2). Therefore this aggregator scales better to large groups of forecasters. On the other hand, problem (7.2)

requires a training set with known outcomes whereas $\boldsymbol{\Sigma}$ can be learned from the forecasts alone. Therefore the two aggregators serve somewhat different purposes and should be considered complementary rather than competitive.

## 7.5 Case Study: Concrete Compressive Strength

Concrete is the most important material in civil engineering. One of its key properties is compressive strength that depends on the water-to-cement ratio but also on several other ingredients. Yeh (1998) illustrated this by statistically predicting compressive strength based on age and seven mixture ingredients. The associated dataset is freely available at the UC Irvine Machine Learning Repository (Lichman, 2013) and consists of $1,030$ observations with the following information:

$$
\begin{array}{lll}
 & Y: & \text{Compressive Strength} \\
\mathcal{M}_F\Big\{ & \mathcal{M}_1\Big\{ & v_1: \text{ Cement (kg in a } m^3 \text{ mixture)} \\
 & & v_2: \text{ Coarse Aggregate (kg in a } m^3 \text{ mixture)} \\
 & & v_3: \text{ Fly Ash (kg in a } m^3 \text{ mixture)} \\
 & & v_4: \text{ Water (kg in a } m^3 \text{ mixture)} \\
 & \mathcal{M}_2\Big\{ & v_5: \text{ Superplasticizer (kg in a } m^3 \text{ mixture)} \\
 & & v_6: \text{ Fine Aggregate (kg in a } m^3 \text{ mixture)} \\
 & & v_7: \text{ Blast Furnace Slag (kg in a } m^3 \text{ mixture)} \\
 & & v_8: \text{ Age (days)}
\end{array}
\Big\}\mathcal{M}_3 \tag{7.5}
$$

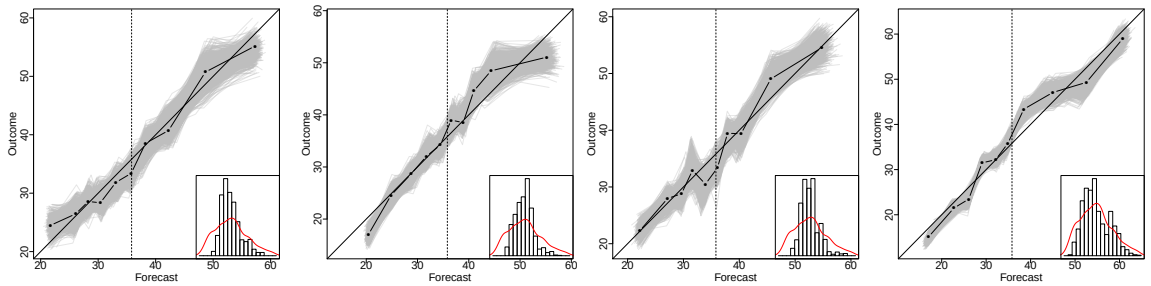
This particular dataset is appropriate for illustrating our results because it is simple yet large enough to allow the computation of reliability diagrams and the individual components of the average quadratic loss.

The individual forecasters are emulated with three linear regression models, $\mathcal{M}_1$, $\mathcal{M}_2$,

and $\mathcal{M}_3$, that predict $Y$ based on different sets of predictors. In particular, model $\mathcal{M}_1$ only uses predictors $v_1, v_2, v_3, v_4$, whereas model $\mathcal{M}_2$ uses the remaining predictors $v_5, v_6, v_7, v_8$. Therefore their predictor sets are non-overlapping. The third model $\mathcal{M}_3$ uses the middle four predictors $v_3, v_4, v_5, v_6$, and hence has significant overlap with the other two models. The results are compared against a linear regression model $\mathcal{M}_F$ that has access to all eight predictors. This is not an aggregator and only represents the extent to which the predictors can explain the outcome $Y$. Therefore it provides interpretation and scale. The predictor sets corresponding to the different models are summarized by the curly braces in (7.5). Overall, this setup can be viewed as a real-valued equivalent of the case study in Ranjan and Gneiting (2010) who aggregate probability forecasts from three different logistic regression models.

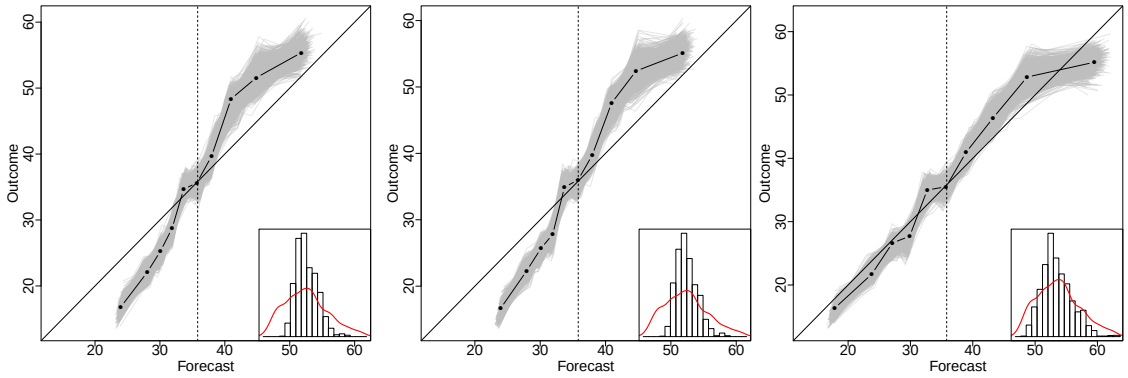
The evaluation is based on a 10-fold cross validation. The models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are first trained on one half of the training set and then used to make predictions for the second half and the entire testing set. Next, the aggregators are trained on the models' predictions over the second half of the training set. Finally, the trained aggregators are tested on the models' predictions over the testing set. Therefore all the following results, apart from the parameter estimates, represent out-of-sample performance. Similarly to Section A.4.3, the evaluation is performed separately under two different information structures: the *No Information Overlap* scenario considers only predictions from models $\mathcal{M}_1$ and $\mathcal{M}_2$, whereas the *High Information Overlap* scenario involves only predictions from models $\mathcal{M}_1$ and $\mathcal{M}_3$.

Figures 7.4, 7.5, and 7.6 present the reliability diagrams of the individual models and the aggregators under no and high information overlap, respectively. Unlike in Section A.4.3, the marginal distribution of $Y$ is not known. Therefore the red curve over the inlined histogram represents the empirical distribution of $Y$. Similarly, the dashed vertical line represents the sample average of the outcomes instead of the marginal mean $\mu_0$. According to




(a) $\mathcal{M}_1$ (b) $\mathcal{M}_2$ (c) $\mathcal{M}_3$ (d) $\mathcal{M}_F$

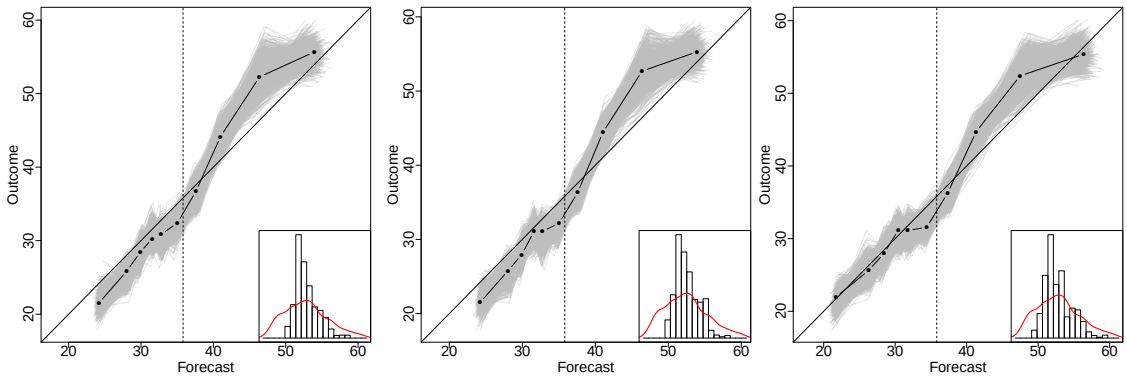
Figure 7.4: Real-World Data. Out-of-sample reliability of the individual models.




(a) $\bar{\mathcal{X}}$ (b) $\mathcal{X}_w$ (c) $\mathcal{X}^*$

Figure 7.5: Real-World Data. Out-of-sample reliability of aggregators under no information overlap.

these plots, the individual forecasts are mostly reliable, except at extremely small or large forecasts. The averaging aggregators $\bar{\mathcal{X}}$ and $\mathcal{X}_w$, on the other hand, are both unreliable and under-confident. Similarly to Section A.4.3 and in accordance with Theorem 7.2.2, this under-confidence decreases as the forecasters' information overlap increases from Figure 7.5 to Figure 7.6. Table 7.3 gives the parameter estimates for $\mathcal{X}_w$ and $\mathcal{X}^*$. These aggregators employ very similar weights. In both information scenarios $\alpha > 1$, suggesting that $\mathcal{X}_w$ is under-confident and should be extremized as it is. Based on Figures 7.5c and 7.6c, the resulting aggregator $\mathcal{X}^*$ is noticeably more reliable and appears to approximate the empirical distribution of $Y$ quite closely. Simply based on visual assessment $\mathcal{X}^*$ performs as well as




(a) $\bar{\mathcal{X}}$ (b) $\mathcal{X}_w$ (c) $\mathcal{X}^*$

Figure 7.6: Real-World Data. Out-of-sample reliability of aggregators under high information overlap.

Table 7.3: Real-World Data. Estimated parameter values.

| Scenario | Forecast | $\mu_0$ | $\alpha$ | $w_1$ | $w_2$ |
|---|---|---|---|---|---|
| No Overlap | $\mathcal{X}_w$ | | | 0.5327 | 0.4673 |
| | $\mathcal{X}^*$ | -36.2051 | 1.6950 | 0.5269 | 0.4731 |
| High Overlap | $\mathcal{X}_w$ | | | 0.5931 | 0.4069 |
| | $\mathcal{X}^*$ | -37.6776 | 1.4382 | 0.5375 | 0.4625 |

$\mathcal{M}_F$ under low information overlap but loses some resolution once overlap is introduced. This makes sense because the models considered in the high information overlap scenario, namely $\mathcal{M}_1$ and $\mathcal{M}_3$ have access only to the first six predictors while $\mathcal{M}_F$ uses all eight predictors and hence should have a higher level of information.

Table 7.4 provides a numerical comparison by presenting the average quadratic loss, its additive components, and the estimated variance $s^2$ for the individual models and the competing aggregators. Given that all aggregators perform better than the individual forecasters, aggregation is generally beneficial. However, there are large performance differences among the aggregators. In particular, the variances of $\bar{\mathcal{X}}$ and $\mathcal{X}_w$ do not exceed that of the individual forecasters', suggesting that neither of them is expanding. Furthermore, they are much less reliable than the individual forecasters. In contrast, $\mathcal{X}^*$ is able to maintain the

Table 7.4: Real-World Data. The average quadratic loss, $L(\mathbf{Y}, \boldsymbol{\mathcal{X}})$ with its three additive components: reliability (REL), resolution (RES), and uncertainty (UNC). The final column, $s^2$ gives the estimated variance of the forecast.

| Scenario | Forecast | $L(\mathbf{Y}, \boldsymbol{\mathcal{X}})$ | REL | RES | UNC | $s^2$ |
|---|---|---|---|---|---|---|
| | $\mathcal{M}_1$ | 187.80 | 9.70 | 100.72 | 278.81 | 82.83 |
| | $\mathcal{M}_2$ | 185.74 | 12.01 | 105.08 | 278.81 | 92.51 |
| | $\mathcal{M}_3$ | 197.03 | 12.81 | 94.59 | 278.81 | 73.27 |
| | $\mathcal{M}_F$ | 110.91 | 9.46 | 177.36 | 278.81 | 157.87 |
| | $\bar{\mathcal{X}}$ | 155.69 | 30.99 | 154.10 | 278.81 | 56.33 |
| No Overlap | $\mathcal{X}_w$ | 156.32 | 31.45 | 153.94 | 278.81 | 56.21 |
| | $\mathcal{X}^*$ | 133.23 | 9.86 | 155.45 | 278.81 | 161.89 |
| | $\bar{\mathcal{X}}$ | 177.45 | 16.77 | 118.13 | 278.81 | 61.92 |
| High Overlap | $\mathcal{X}_w$ | 176.59 | 14.37 | 116.59 | 278.81 | 63.32 |
| | $\mathcal{X}^*$ | 169.92 | 8.20 | 117.09 | 278.81 | 128.69 |

forecasters' level of reliability. Even though this aggregator is expanding, it is less resolute and has a lower variance than $\mathcal{M}_F$ under high information overlap. This can be expected because in the high information overlap scenario $\mathcal{X}^*$ has access only to a subset of the information that $\mathcal{M}_F$ uses. Under no information overlap, all the predictors are used by the individual forecasters, but this does not mean that this information is actually revealed to $\mathcal{X}^*$ through the reported forecasts.

## 7.6 Summary and Discussion

This paper discussed forecast aggregation under a general probability model, called the partial information framework. The forecasts and outcomes were assumed to have a joint distribution but no restrictions were placed on their dependence structure. The analysis led to an enumeration (Theorem 7.2.1) of several properties of optimal aggregation. Even though the optimal aggregator is typically intractable in practice, its properties provide guidance for developing and understanding other aggregators that are more feasible in

practice. In this paper these properties shed light on the class of weighted averages of any type of univariate forecasts. Even though these averages are marginally consistent, they fail to satisfy two of the optimality properties, namely reliability and variance expansion (Theorem 7.2.2). As a result, they are under-confident in a sense that they are overly close to the marginal mean. This shortcoming can be naturally alleviated by extremizing, that is, by shifting the weighted average further away from the marginal mean. Section 7.3 introduced a simple linear procedure (Equation 7.1) that extremizes the weighted average of real-valued forecasts and maintains marginal consistency. This procedure and the theoretical results were illustrated on synthetic (Section A.4.3) and real-world data (Section 7.5). In both cases the optimally weighted average was shown to be both unreliable and under-confident, especially when the forecasters used very different sets of information. Fortunately, extremization was able to largely correct these drawbacks and provide transformed aggregates that were both reliable and more resolute.

Forecast aggregation literature by and large agrees that the goal is to collect and combine information from different forecasters (see, e.g., Dawid et al. 1995; Armstrong 2001; Forlines et al. 2012). At the same time aggregation continues to be performed via weighted averaging or perhaps some other measure of central tendency, such as the median (Levins, 1966; Armstrong, 2001; Lobo and Yao, 2010). Section 7.2.2 explained that these popular techniques do not behave like aggregators of information. Instead, they are designed to reduce measurement error which is philosophically very different from information diversity (Satopää et al., 2016). Therefore some details of their workings seem to have been misunderstood. Unfortunately, it is unlikely that this paper will prevent aggregation with measures of central tendency all together. However, it is hoped that our contributions will at least prompt interest and provide direction in discovering alternative aggregation techniques.

This paper illustrated that good information aggregation can arise from a simple linear

transformation that extremizes the weighted average. Of course, under a large number of prediction problems, a non-linear extremizing function can lead to further improvements in aggregation. The linear function, however, is a simple and natural starting point that suffices for illustrating the benefits of extremizing. Is extremizing then guaranteed to be beneficial in every prediction task? Probably not. Therefore, for the sake of applications, it is important to discuss conditions under which extremizing is likely to improve the commonly used aggregators. Item iv) of Theorem 7.2.2 and the empirical results in Sections A.4.3 and 7.5 suggest that extremizing is likely to be more beneficial under no or low information overlap. This aligns with Satopää et al. (2015) who use the Gaussian partial information model to show empirically that extremizing probability forecasts becomes more important a) as the amount of the forecasters' combined information increases, and b) as the forecasters' information sets become more diverse. This means that, for instance, the average forecast of team members working in close collaboration require little extremizing whereas forecasts coming from widely different sources must be heavily extremized.

Unfortunately, the amount and direction of extremization depends on a training set with known outcomes. Such a training set may not always be available. In the most extreme case the decision-maker may have only a set of forecasts of a single unknown outcome. How should the forecasts be aggregated in such a low-data setting? The results in this paper suggest that any type of weighted average (or some other measure of central tendency) is a poor choice. A better alternative was discussed by Satopää et al. (2015). They assume that the forecasters' covariance matrix is compound symmetric and then aggregate the probability forecasts with the optimal aggregator under the corresponding Gaussian partial information model. Developing more general aggregators that place less constraints on the joint dependence structure while satisfying at least two of the optimality properties of Theorem 7.2.1 is certainly an interesting future research direction. The first step is to develop a simple aggregator that is both marginally consistent and expanding. Finding an

aggregator that maintains forecasters' reliability seems more difficult.

## 7.7 Acknowledgements

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

# 8

# Conclusion and Future Work

This dissertation proposed an alternative to a core statistical concept, namely measurement error, and therefore led to a new modeling paradigm that is fundamentally different from classical statistics. Given that the contribution is at the root of statistical theory, the basic theory of information diversity suggests a range of theoretical and applied projects in statistics and other related fields. The following enumeration illustrates some specific projects.

1. So far, the Gaussian model has been mostly applied to forecasters predicting multiple related outcomes. Therefore a natural next step is to develop an aggregator for a set of forecasts of a single outcome. One idea is to first derive the revealed aggregator based on the Gaussian model and then integrate out any unknown parameters with respect to their posterior distribution. Chapter 6 introduced such an aggregator for two probability forecasts under some restrictive assumptions on their information structure. An extension to $N$ forecasters with less restrictions on the information structure is certainly needed. One idea is to treat the forecasters as exchangeable and hence model their information structure with a compound symmetric covariance matrix. A slightly more involved version would treat the forecasters as partially exchangeable and model their information structure with a covariance matrix consisting of compound symmetric blocks.

2. Satopää et al. (2016) estimate the information structure based on multiple predictions per forecaster. The rest of the model parameters are estimated separately, leading to an overall procedure with multiple steps. A more principled approach would estimate all parameters jointly. Such joint estimation tasks are often easier within the Bayesian paradigm where many numerical tools exists for estimating complex hierarchical models. The only non-standard challenge here is to develop a prior distribution for the information structure – a problem that seems interesting in itself.

3. Information diversity is not only relevant in statistical estimation but also offers new directions in other areas of statistics. For instance, information diversity can be applied to testing hypotheses about a binary outcome. Such testing procedures can be expected to yield higher power when information diversity is the dominant source of data variation.

4. Most aggregation procedures use one estimate per forecaster, even though it is common for experts to update their beliefs over time. Motivated by this, Chapter 3 considered a dynamic context in which experts can update their beliefs at random intervals. The aggregator therein, however, is an empirical procedure that requires a training set of multiple prediction problems with known outcomes. Therefore, an interesting future project is to use martingale theory within the partial information framework and construct a time-series aggregator that can operate directly on the forecasts.

5. Information diversity can be also applied to model predictions. This suggests many contributions in machine learning. For instance, semi-supervised learning refers to a setup with a large amount of data of which only a relatively few have been labeled. While the commonly used ensemble techniques, such as stacking or Bayesian averaging, require labeled observations to combine the individual models' predictions, partial information aggregators do not. This suggests an opportunity for improved

aggregation efficiency.

6. In many applications all the forecasts may not be available at any given moment but instead arrive in a sequential manner. In such settings, re-computing the information structure every time a new prediction arrives can be computationally demanding or infeasible. This motivates the development of a sequential aggregation approach. Such a procedure could keep track of the current consensus and the information it is based upon. Any new forecast is then aggregated directly into this running consensus. Therefore aggregation always involves only two forecasts: the current consensus and the new forecast. This significantly reduces the dimension of the problem and hence can facilitate estimation techniques that are both theoretically sound and more efficient.

7. Under the Gaussian model a forecast that is further away from the marginal mean is considered more informed. Of course, the more informed forecasters typically have a higher impact on the final aggregate. This suggests that an uncalibrated forecaster who gives overly extreme predictions can interfere with proper aggregation. One solution to such ill-behaved forecasts is to develop a more robust version of the revealed aggregator. This can be achieved by explicitly controlling the amount of information that any forecaster can have, or alternatively by deriving the revealed aggregator under a distribution with heavy tails, such as the multivariate $t$-distribution.

8. Even though a large literature attests to the benefits of collective problem solving and crowd wisdom, it is generally less clear how groups should be formed, that is, what kind of individuals brought together yield optimal results. Page (2008) argues in favor of diversity: more diverse groups yield better results. This general statement, however, requires some further analysis. In particular, it is unlikely that diversity helps in every forecasting application, and in all those applications where diversity

helps, it is unlikely that the type of beneficial diversity remains stable. For instance, in predicting the success of a new blockbuster, it is important to have both men and women in the group. On the other hand, men are likely to be less helpful in predicting the demand of a new designer dress. The importance of diversity could be investigated by developing a partial information model that allows the covariance matrix to depend on the forecasters' personal characteristics such as sex, education, ethnicity, age, and so on. Such a model could pinpoint the beneficial type of diversity in any given application and then use this knowledge to automatically choose the right group of individuals to contribute in the final the aggregate forecast.

9. One theoretical direction is to show that no measure of central tendency collects information. Proving such a statement, of course, would need a general characterization of measures of central tendency. One option is to analyze aggregators that cannot leave the convex hull of the individual forecasts. This result would naturally lead to a discussion about the functional form of aggregators that do preserve good properties, such as calibration, of the individual forecasts. Given that information manifests itself in terms of variance, it is unlikely that an aggregator that depends only on the first moment can preserve calibration. This all, however, must be investigated and made precise.

10. Another theoretical direction involves estimation of information overlap. First, given that $\mathrm{Cov}\,(X_i, X_j)$ can be slightly negative, it cannot be a precise measure of information overlap. This overlap is given by $\mathcal{F}_{ij} = \mathcal{F}_i \cap \mathcal{F}_j$. Given that $\mathcal{F}_{ij}$ is a $\sigma$-field, there exists a forecast $X_{ij}$ such that $X_{ij} = \mathbb{E}(Y|\mathcal{F}_{ij})$. The variance of this forecast, namely $\mathrm{Var}\,(X_{ij})$ quantifies the information overlap $\mathcal{F}_{ij}$. How can this partial variance be captured in practice, and how does it relate to $\mathrm{Cov}\,(X_i, X_j)$? Second, the Gaussian model only incorporates pairwise information overlaps. DeGroot and Mortera (1991) show that considering pairwise overlaps is enough for weighted av-

eraging. Intuitively, a similar result should not hold under the partial information framework because the revealed aggregator aims to use the group's combined information. The amount of this information cannot be determined only based on the sizes of the individual information sets and their pairwise overlaps. If higher order overlaps turn out to matter, the Gaussian model should be replaced by a new specification that can incorporate such overlaps. This is likely to involve generalizing the usual covariance to a new measure of how much any number of random variables change together.

# A Appendices

## A.1 Supplement for Chapter 2

## Appendix

Unfortunately, the full real-world dataset is not accessible for the public at the moment. We have, however, requested a permission to publish the data online in the near future. For the time being, we have included the following table that shows a complete list of the 69 problems in our dataset. For each problem, six summary statistics have been provided:

$\hat{p}_G$ = Our aggregate estimate based on the forecasts made within the first three days. The bias term, $a$, was estimated with $\hat{a}_{MLE}$.

$\bar{p}$ = Sample average of the forecasts made within the first three days.

$s_p$ = Sample standard deviation of the forecasts made within the first three days.

$N$ = Number of forecasts made within the first three days.

$T$ = Number of days that the problem was open.

$Z$ = Indicator on whether the event happened ($Z = 1$) or did

not happen ($Z = 0$).

Even though this paper does not focus on dynamic data, we report the time-frame of each problem because this is somewhat indicative of the uncertainty and difficulty of the problem.

| Question Text | $\hat{p}_G$ | $\bar{p}$ | $s_p$ | $N$ | $T$ | $Z$ |
|---|---|---|---|---|---|---|
| Will the Six-Party talks (among the US, North Korea, South Korea, Russia, China, and Japan) formally resume in 2011? | 0.04 | 0.27 | 0.22 | 102 | 123 | 0 |
| Will Serbia be officially granted EU candidacy by 31 December 2011? | 0.03 | 0.27 | 0.24 | 96 | 124 | 0 |
| Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011? | 0.01 | 0.23 | 0.27 | 128 | 29 | 0 |
| Will Daniel Ortega win another term as President of Nicaragua during the late 2011 elections? | 0.78 | 0.61 | 0.19 | 109 | 65 | 1 |
| Will Italy restructure or default on its debt by 31 December 2011? | 0.28 | 0.44 | 0.26 | 86 | 124 | 0 |
| By 31 December 2011, will the World Trade Organization General Council or Ministerial Conference approve the 'accession package' for WTO membership for Russia? | 0.43 | 0.48 | 0.21 | 98 | 106 | 1 |
| Will the 30 Sept 2011 "last" PPB for Nov 2011 Brent Crude oil futures exceed $115? | 0.23 | 0.40 | 0.21 | 302 | 23 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will the Nikkei 225 index finish trading at or above 9,500 on 30 September 2011? | 0.06 | 0.29 | 0.21 | 290 | 22 | 0 |
| Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 October 2011? | 0.03 | 0.24 | 0.20 | 333 | 23 | 0 |
| Will the London Gold Market Fixing price of gold (USD per ounce) exceed $1850 on 30 September 2011 (10am ET)? | 0.78 | 0.60 | 0.22 | 269 | 23 | 0 |
| Will Israel's ambassador be formally invited to return to Turkey by 30 September 2011? | 0.02 | 0.22 | 0.19 | 334 | 23 | 0 |
| Will PM Donald Tusk's Civic Platform Party win more seats than any other party in the October 2011 Polish parliamentary elections? | 0.80 | 0.61 | 0.19 | 281 | 31 | 1 |
| Will Robert Mugabe cease to be President of Zimbabwe by 30 September 2011? | 0.01 | 0.16 | 0.20 | 358 | 23 | 0 |
| Will Muqtada al-Sadr formally withdraw support for the current Iraqi government of Nouri al-Maliki by 30 September 2011? | 0.08 | 0.30 | 0.19 | 282 | 23 | 0 |
| Will peace talks between Israel and Palestine formally resume at some point between 3 October 2011 and 1 November 2011? | 0.02 | 0.23 | 0.21 | 309 | 28 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will the expansion of the European bailout fund be ratified by all 17 Eurozone nations before 1 November 2011? | 0.64 | 0.55 | 0.26 | 395 | 9 | 1 |
| Will the South African government grant the Dalai Lama a visa before 7 October 2011? | 0.03 | 0.28 | 0.25 | 647 | 2 | 0 |
| Will former Ukrainian Prime Minister Yulia Tymoshenko be found guilty on any charges in a Ukrainian court before 1 November 2011? | 0.50 | 0.51 | 0.20 | 364 | 6 | 1 |
| Will Abdoulaye Wade win re-election as President of Senegal? | 0.79 | 0.62 | 0.16 | 200 | 173 | 0 |
| Will the Freedom and Justice Party win at least 20 percent of the seats in the first People's Assembly (Majlis al-Sha'b) election in post-Mubarak Egypt? | 0.86 | 0.65 | 0.19 | 207 | 108 | 1 |
| Will Joseph Kabila remain president of the Democratic Republic of the Congo through 31 January 2012? | 0.93 | 0.72 | 0.16 | 166 | 119 | 1 |
| Will Moody's issue a new downgrade of the sovereign debt rating of the Government of Greece between 3 October 2011 and 30 November 2011? | 0.83 | 0.64 | 0.22 | 203 | 57 | 0 |
| Will the UN Security Council pass a measure/resolution concerning Syria in October 2011? | 0.11 | 0.35 | 0.24 | 231 | 27 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will the U.S. Congress pass a joint resolution of disapproval in October 2011 concerning the proposed $5+ billion F-16 fleet upgrade deal with Taiwan? | 0.02 | 0.23 | 0.22 | 297 | 17 | 0 |
| Will the Japanese government formally announce the decision to buy at least 40 new jet fighters by 30 November 2011? | 0.30 | 0.44 | 0.20 | 193 | 57 | 0 |
| Will the Tunisian Ennahda party officially announce the formation of an interim coalition government by 15 November 2011? | 0.70 | 0.57 | 0.23 | 508 | 7 | 0 |
| Will Japan officially become a member of the Trans-Pacific Partnership before 1 March 2012? | 0.47 | 0.49 | 0.22 | 150 | 113 | 0 |
| Will the United Nations Security Council pass a new resolution concerning Iran by 1 April 2012? | 0.59 | 0.53 | 0.27 | 193 | 145 | 0 |
| Will Hamad bin Isa al-Khalifa remain King of Bahrain through 31 January 2012? | 0.99 | 0.82 | 0.17 | 163 | 84 | 1 |
| Will Bashar al-Assad remain President of Syria through 31 January 2012? | 0.95 | 0.71 | 0.24 | 143 | 84 | 1 |
| Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 January 2012? | 1.00 | 0.83 | 0.22 | 523 | 4 | 1 |
| Will Lucas Papademos be the next Prime Minister of Greece? | 0.94 | 0.70 | 0.24 | 388 | 2 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will Lucas Papademos resign, lose re-election/confidence vote, or vacate the office of Prime Minister of Greece before 1 March 2012? | 0.17 | 0.38 | 0.24 | 231 | 79 | 0 |
| Will the United Kingdom's Tehran embassy officially reopen by 29 February 2012? | 0.02 | 0.21 | 0.20 | 237 | 79 | 0 |
| Will a trial for Saif al-Islam Gaddafi begin in any venue by 31 March 2012? | 0.41 | 0.48 | 0.26 | 215 | 110 | 0 |
| Will S&P downgrade the AAA long-term credit rating of the European Financial Stability Facility (EFSF) by 30 March 2012? | 0.69 | 0.57 | 0.22 | 259 | 33 | 1 |
| Will North Korea successfully detonate a nuclear weapon, either atmospherically, underground, or underwater, between 9 January 2012 and 1 April 2012? | 0.02 | 0.22 | 0.22 | 215 | 83 | 0 |
| By 1 April 2012, will Egypt officially announce its withdrawal from its 1979 peace treaty with Israel? | 0.01 | 0.18 | 0.19 | 227 | 83 | 0 |
| Will Kim Jong-un attend an official, in-person meeting with any G8 head of government before 1 April 2012? | 0.02 | 0.21 | 0.22 | 238 | 82 | 0 |
| Will Christian Wulff resign or vacate the office of President of Germany before 1 April 2012? | 0.16 | 0.37 | 0.23 | 241 | 38 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will the daily Europe Brent Crude FOB spot price per barrel be greater than or equal to $150 before 3 April 2012? | 0.04 | 0.27 | 0.23 | 206 | 84 | 0 |
| Will the Taliban begin official in-person negotiations with either the US or Afghan government by 1 April 2012? | 0.08 | 0.32 | 0.24 | 184 | 69 | 0 |
| Will Yousaf Raza Gillani resign, lose confidence vote, or vacate the office of Prime Minister of Pakistan before 1 April 2012? | 0.18 | 0.38 | 0.22 | 146 | 68 | 0 |
| Will Yemen's next presidential election commence before 1 April 2012? | 0.34 | 0.46 | 0.25 | 222 | 28 | 1 |
| Will Traian Basescu resign, lose referendum vote, or vacate the office of President of Romania before 1 April 2012? | 0.08 | 0.31 | 0.21 | 149 | 68 | 0 |
| Will the UN Security Council pass a new measure/resolution directly concerning Syria between 23 January 2012 and 31 March 2012? | 0.39 | 0.47 | 0.26 | 156 | 68 | 0 |
| Before 1 April 2012, will South Korea officially announce a policy of reducing Iranian oil imports in 2012? | 0.46 | 0.49 | 0.24 | 170 | 68 | 0 |
| Will Israel release Palestinian politician Aziz Duwaik from prison before 1 March 2012? | 0.05 | 0.29 | 0.23 | 210 | 37 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will Iran and the U.S. commence official nuclear program talks before 1 April 2012? | 0.01 | 0.17 | 0.20 | 225 | 61 | 0 |
| Will Serbia be officially granted EU candidacy before 1 April 2012? | 0.07 | 0.30 | 0.23 | 253 | 31 | 1 |
| Will the IMF officially announce before 1 April 2012 that an agreement has been reached to lend Hungary an additional 15+ Billion Euros? | 0.51 | 0.51 | 0.23 | 177 | 61 | 0 |
| Will Libyan government forces regain control of the city of Bani Walid before 6 February 2012? | 0.18 | 0.38 | 0.24 | 500 | 6 | 0 |
| Will a run-off be required in the 2012 Russian presidential election? | 0.04 | 0.29 | 0.25 | 277 | 34 | 0 |
| Will the Iraqi government officially announce before 1 April 2012 that it has dropped all criminal charges against its VP Tareq al-Hashemi? | 0.04 | 0.27 | 0.22 | 200 | 61 | 0 |
| Will Egypt officially announce by 15 February 2012 that it is lifting its travel ban on Americans currently in Egypt? | 0.44 | 0.50 | 0.25 | 321 | 16 | 0 |
| Will a Japanese whaling ship enter Australia's territorial waters between 7 February 2012 and 10 April 2012? | 0.17 | 0.37 | 0.27 | 213 | 63 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will William Ruto cease to be a candidate for President of Kenya before 10 April 2012? | 0.16 | 0.37 | 0.25 | 192 | 62 | 0 |
| Will Marine LePen cease to be a candidate for President of France before 10 April 2012? | 0.04 | 0.26 | 0.22 | 214 | 62 | 0 |
| Between 21 February 2012 and 1 April 2012, will the UN Security Council announce any reduction of its peacekeeping force in Haiti? | 0.20 | 0.41 | 0.25 | 168 | 40 | 0 |
| Will Mohamed Waheed Hussain Manik resign or otherwise vacate the office of President of Maldives before 10 April 2012? | 0.08 | 0.32 | 0.23 | 155 | 48 | 0 |
| Will Japan commence parliamentary elections before 1 April 2012? | 0.04 | 0.28 | 0.23 | 182 | 39 | 0 |
| Before 13 April 2012, will the Turkish government officially announce that the Turkish ambassador to France has been recalled? | 0.06 | 0.29 | 0.22 | 143 | 51 | 0 |
| Will Standard and Poor's downgrade Japan's Foreign Long Term credit rating at any point between 21 February 2012 and 1 April 2012? | 0.08 | 0.32 | 0.25 | 172 | 40 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Will Myanmar release at least 100 more political prisoners between 21 February 2012 and 1 April 2012? | 0.55 | 0.52 | 0.25 | 170 | 40 | 0 |
| Will a civil war break out in Syria between 21 February 2012 and 1 April 2012? | 0.54 | 0.51 | 0.24 | 191 | 40 | 0 |
| Will Tunisia officially announce an extension of its current state of emergency before 1 April 2012? | 0.77 | 0.61 | 0.24 | 198 | 26 | 1 |
| Before 1 April 2012, will Al-Saadi Gaddafi be extradited to Libya? | 0.02 | 0.22 | 0.20 | 225 | 26 | 0 |
| Before 1 April 2012, will the Sudan and South Sudan governments officially announce an agreement on oil transit fees? | 0.04 | 0.27 | 0.23 | 202 | 26 | 0 |
| Will Yemeni government forces regain control of the towns of Jaar and Zinjibar from Al-Qaida in the Arabian Peninsula (AQAP) before 1 April 2012? | 0.04 | 0.28 | 0.24 | 192 | 26 | 0 |

Figure A.1 summarizes the data by giving a scatterplot of $\hat{p}_G(\hat{a}_{MLE})$ against $\bar{p}$. Notice how the points are above (or below) the 45-degree dashed line when $\hat{p}_G(\hat{a}_{MLE})$ is less (or more) than 0.5. This implies that $\hat{p}_G(\hat{a}_{MLE})$ is a much sharper aggregator than $\bar{p}$.

## A.2 Supplement for Chapter 3

This supplementary material accompanies the paper "Probability Aggregation in Time-Series: Dynamic Hierarchical Modeling of Sparse Expert Beliefs". It provides a technical description of the sampling step of the SAC-algorithm.

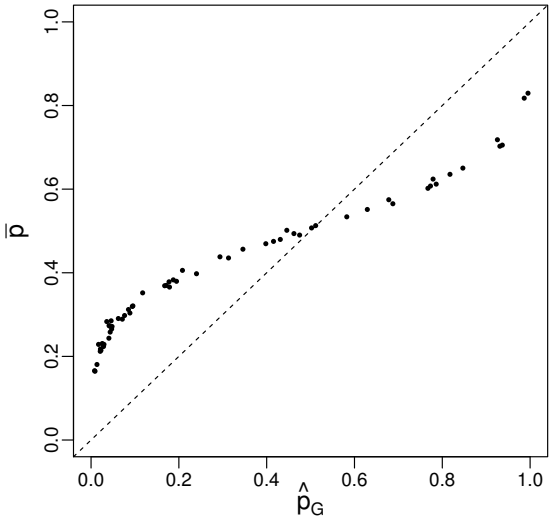



Figure A.1: A summarizing comparison of the aggregators $\hat{p}_G(\hat{a}_{MLE})$ and $\bar{p}$.

# Technical Details of the Sampling Step

The Gibbs sampler (Geman and Geman (1984)) iteratively samples all the unknown parameters from their full-conditional posterior distributions one block of parameters at a time. Given that this is performed under the constraint $b_3 = 1$ to ensure model identifiability, the constrained parameter estimates should be denoted with a trailing $(1)$ to maintain consistency with earlier notation. For instance, the constrained estimate of $\gamma_k$ should be denoted by $\hat{\gamma}_k(1)$ while the unconstrained estimate is denoted by $\hat{\gamma}_k$. For the sake of clarity, however, the constraint suffix is omitted in this section. Nonetheless, it is important to keep in mind that all the estimates in this section are constrained.

## A.2.1 Sample $X_{t,k}$

The hidden states are sampled via the *Forward-Filtering-Backward-Sampling* (FFBS) algorithm that first predicts the hidden states using a Kalman Filter and then performs a backward sampling procedure that treats these predicted states as additional observations

(see, e.g., Carter and Kohn (1994); Migon et al. (2005) for details on FFBS). More specifically, the first part, namely the Kalman Filter, is deterministic and consists of a predict and an update step. Given all the other parameters except the hidden states, the predict step for the $k$th question is

$$X_{t|t-1,k} = \gamma_k X_{t-1|t-1,k}$$

$$P_{t|t-1,k} = \gamma_k^2 P_{t-1|t-1,k} + \tau_k^2,$$

where the initial values, $X_{0|0,k}$ and $P_{0|0,k}$, are equal to $0$ and $1$, respectively.

**Algorithm 2** The update step of the FFBS algorithm. $N_{t,k}$ denotes the number of forecasts made at time $t$ for question $k$. The subindex $j(i)$ denotes the $i$th expert's self-assessed expertise group.

**for** $i = 1, 2, \ldots, N_{t,k}$ **do**
  $e_{i,t,k} = Y_{i,t,k} - b_{j(i)} X_{t|t-1,k}$
  $S_{i,t,k} = \sigma_k^2 + b_{j(i)}^2 P_{t|t-1,k}$
  $K_{i,t,k} = P_{t|t-1,k} b_{j(i)} S_{i,t,k}^{-1}$
  $X_{t|t,k} = X_{t|t-1,k} + K_{i,t,k} e_{i,t,k}$
  $P_{t|t,k} = (1 - K_{i,t,k} b_{j(i)}) P_{t|t-1,k}$
  **if** $i \neq N_{t,k}$ **then**
    $X_{t|t-1,k} = X_{t|t,k}$
    $P_{t|t-1,k} = P_{t|t,k}$
  **end if**
**end for**

The update step is given by Algorithm 2. The update is repeated sequentially for each observation $Y_{i,t,k}$ given at time $t$. For each such repetition, the previous posterior values, $X_{t|t,k}$ and $P_{t|t,k}$, are considered as the new prior values, $X_{t|t-1,k}$ and $P_{t|t-1,k}$. If the observation $\boldsymbol{Y}_{t,k}$ is completely missing at time $t$, the update step is skipped and

$$X_{t|t,k} = X_{t|t-1,k}$$

$$P_{t|t,k} = P_{t|t-1,k}$$

After running the Kalman Filter up to the final time point at $t = T_k$, the final hidden state is sampled from $X_{T_k,k} \sim \mathcal{N}(X_{T_k|T_k,k}, P_{T_k|T_k,k})$. The remaining states are obtained via the backward sampling that is performed in reverse from

$$X_{t-1,k} \sim \mathcal{N}\left(V\left(\frac{\gamma_k X_{t,k}}{\tau_k^2} + \frac{X_{t|t,k}}{P_{t|t,k}}\right), V\right),$$

where

$$V = \left(\frac{\gamma_k^2}{\tau_k^2} + \frac{1}{P_{t|t,k}}\right)^{-1}$$

This can be viewed as backward updating that considers the Kalman Filter estimates as additional observations at each given time point.

### A.2.2 Sample $\boldsymbol{b}$ and $\sigma_k^2$

First, vectorize all the response vectors $\boldsymbol{Y}_{t,k}$ into a single vector denoted

$$\boldsymbol{Y}_k = \left[\boldsymbol{Y}_{1,k}^T, \ldots, \boldsymbol{Y}_{T_k,k}^T\right]^T.$$

Given that each $\boldsymbol{Y}_{t,k}$ is matched with $X_{t,k}$ via the time index $t$, we can form a $|\boldsymbol{Y}_k| \times J$ design-matrix by letting

$$\boldsymbol{X}_k = \left[(\boldsymbol{M}_k X_{1,k})^T, \ldots, (\boldsymbol{M}_k X_{T_k,k})^T\right]^T$$

Given that the goal is to borrow strength across questions by assuming a common bias vector $\boldsymbol{b}$, the parameter values must be estimated in parallel for each question such that the matrices $\boldsymbol{X}_k$ can be further concatenated into $\boldsymbol{X} = [\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T]^T$ during every iteration. Similarly, $\boldsymbol{Y}_k$ must be further vectorized into a vector $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_K^T]^T$. The question-

specific variance terms are taken into account by letting $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2 \mathbf{1}_{1\times T_1}, \ldots, \sigma_K^2 \mathbf{1}_{1\times T_K})$. After adopting the non-informative prior $p(\boldsymbol{b}, \sigma_k^2 | \boldsymbol{X}_k) \propto \sigma_k^{-2}$ for each $k = 1, \ldots, K$, the bias vector is sampled from

$$\boldsymbol{b}|\ldots \;\sim\; \mathcal{N}_J\left((\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\right) \tag{A.1}$$

Given that the covariance matrix in Equation (A.1) is diagonal, the identifiability constraint can be enforced after sampling a new value of $\boldsymbol{b}$ by letting $b_3 = 1$. The variance parameters are then sampled from

$$\sigma_k^2|\ldots \;\sim\; \text{Inv-}\chi^2\left(|\boldsymbol{Y}_k| - J, \frac{1}{|\boldsymbol{Y}_k| - J}(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{b})^T(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{b})\right),$$

where the distribution is a scaled inverse-$\chi^2$ (see, e.g., Gelman et al. (2003)). Given that the experts are not required to give a new forecast at every time unit, the design matrices must be trimmed accordingly such that their dimensions match up with the dimensions of the observed matrices.

### A.2.3 Sample $\gamma_k$ and $\tau_k^2$

The parameters of the hidden process are estimated via a regression setup. More specifically, after adopting the non-informative prior $p(\gamma_k, \tau_k^2 | \boldsymbol{X}_k) \propto \tau_k^{-2}$, the parameter values are sampled from

$$\begin{aligned}
\gamma_k|\ldots \;&\sim\; \mathcal{N}\left(\frac{\sum_{t=2}^{T_k} X_{t,k}X_{t-1,k}}{\sum_{t=1}^{T_k-1} X_{t,k}^2}, \frac{\tau_k^2}{\sum_{t=1}^{T_k-1} X_{t,k}^2}\right) \\
\tau_k^2|\ldots \;&\sim\; \text{Inv-}\chi^2\left(T_k - 1, \frac{1}{T_k - 1}\sum_{t=2}^{T_k}(X_{t,k} - \gamma_k X_{t-1,k})^2\right),
\end{aligned}$$

where the final distribution is a scaled inverse-$\chi^2$ (see, e.g., Gelman et al. (2003)).

## A.3 Supplement for Chapter 4

# Appendix A: Proofs and Derivations

### A.3.1 Proof of Proposition 4.3.3

Denote the set of all coherent information structures with $\mathcal{Q}_N$. Consider $\mathbf{\Sigma}_{22} \in \mathcal{Q}_N$ and its associated Borel sets $\{B_i : i = 1, \ldots, N\}$. Given that $\mathbf{\Sigma}_{22}$ is coherent, its information can be represented in a diagram such as the one given by Figure 1 in the main manuscript. Keeping the diagram representation in mind, partition the unit interval $S$ into $2^N$ disjoint parts $C_{\mathbf{v}} := \cap_{i \in \mathbf{v}} B_i \setminus \cup_{i \notin \mathbf{v}} B_i$, where $\mathbf{v} \subseteq \{1, \ldots, N\}$ denotes a subset of forecasters and each $C_{\mathbf{v}}$ represents information used only by the forecasters in $\mathbf{v}$. Given that $\sum_{\mathbf{v}} |C_{\mathbf{v}}| = 1$, it is possible to establish a linear function $L$ from the probability simplex

$$\begin{aligned} \Delta_N &:= \text{conv}\{\boldsymbol{e}_{\mathbf{v}} : \mathbf{v} \subseteq \{1, \ldots, N\}\} \\ &= \left\{\boldsymbol{z} \in \mathbb{R}^{2^N} : \boldsymbol{z} \geq \mathbf{0}, \mathbf{1}'\boldsymbol{z} = 1\right\} \end{aligned}$$

to the space of coherent information structures $\mathcal{Q}_N$. In particular, the linear function $L : \boldsymbol{z} \in \Delta_N \to \mathbf{\Sigma}_{22} \in \mathcal{Q}_N$ is defined such that $\rho_{ij} = \sum_{\{i,j\} \subseteq \mathbf{v}} z_{\mathbf{v}}$ and $\delta_i = \sum_{i \in \mathbf{v}} z_{\mathbf{v}}$. Therefore $L(\Delta_N) = \mathcal{Q}_N$. Furthermore, given that $\Delta_N$ is a convex polytope,

$$\begin{aligned} L(\Delta_N) &= \text{conv}\{L(\boldsymbol{e}_{\mathbf{v}}) : \mathbf{v} \subseteq \{1, \ldots, N\}\} \qquad \text{(A.2)} \\ &= \text{conv}\left\{\boldsymbol{x}\boldsymbol{x}' : \boldsymbol{x} \in \{0,1\}^N\right\} \\ &= \text{COR}(N), \end{aligned}$$

which establishes $\text{COR}(N) = \mathcal{Q}_N$. Equality (A.2) follows from the basic properties of convex polytopes (see, e.g., McMullen and Shephard 1971, pp. 16). Each $\mathbf{\Sigma}_{22} \in \text{COR}(N)$ has $\frac{N(N+1)}{2} = \binom{n+1}{2}$ parameters and therefore exists in $\binom{n+1}{2}$ dimensions. □

### A.3.2 Proof of Proposition 4.4.1

The proposition is proved by showing $\mathbb{E}(\mathbf{1}_A|\{X_{B_i}\}_{i=1}^N, X_{B'}) = \mathbb{E}(\mathbf{1}_A|X_{B'})$. First, append $X_{B'}$ to the multivariate Gaussian distribution (2) of the main manuscript:

$$\begin{pmatrix} X_S \\ X_{B'} \\ X_{B_1} \\ X_{B_2} \\ \vdots \\ X_{B_N} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Omega_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{pmatrix} = \left(\begin{array}{c|ccccc} 1 & \delta' & \delta_1 & \delta_2 & \dots & \delta_N \\ \hline \delta' & \delta' & \delta_1 & \delta_2 & \dots & \delta_N \\ \delta_1 & \delta_1 & \delta_1 & \rho_{1,2} & \dots & \rho_{1,N} \\ \delta_2 & \delta_2 & \rho_{2,1} & \delta_2 & \dots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \delta_N & \rho_{N,1} & \rho_{N,2} & \dots & \delta_N \end{array}\right)\right).$$

Denote $\mathbf{X}_\Omega = (X_{B'}, X_{B_1}, \dots, X_{B_N})'$. If $\mathbf{e}_1$ is the first standard basis vector of length $N+1$ and the above multivariate Gaussian distribution is non-degenerate, then $\mathbf{\Omega}_{21} = \mathbf{e}_1'\mathbf{\Omega}_{22} \Leftrightarrow \mathbf{\Omega}_{21}\mathbf{\Omega}_{22}^{-1} = \mathbf{e}_1'$. This identity together with the well-known results of the conditional Gaussian distributions (see, e.g., Ravishanker and Dey 2001, Result 5.2.10) give

$$\begin{aligned} \mathbb{E}(\mathbf{1}_A|\{X_{B_i}\}_{i=1}^N, X_{B'}) &= \Phi\left(\frac{\mathbf{\Omega}_{12}\mathbf{\Omega}_{21}^{-1}\mathbf{X}_\Omega}{\sqrt{1-\mathbf{\Omega}_{12}\mathbf{\Omega}_{21}^{-1}\mathbf{\Omega}_{21}}}\right) \\ &= \Phi\left(\frac{\mathbf{e}_1'\mathbf{X}_\Omega}{\sqrt{1-\mathbf{e}_1'\mathbf{\Omega}_{21}}}\right) \\ &= \Phi\left(\frac{X_{B'}}{\sqrt{1-\delta'}}\right) \\ &= \mathbb{E}(\mathbf{1}_A|X_{B'}) \end{aligned}$$

□

### A.3.3 Proof of Proposition 4.4.2

Given that

$$P' \sim \mathcal{N}\left(0, \sigma_1^2 := \frac{\delta'}{1-\delta'}\right)$$
$$\frac{1}{N}\sum_{i=1}^{N} P_i \sim \mathcal{N}\left(0, \sigma_2^2 := \frac{1}{N^2}\left\{\sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}\right\}\right),$$

the amount of extremizing $\alpha$ is a ratio of two correlated Gaussian random variables. The Pearson product-moment correlation coefficient for them is

$$\kappa = \frac{\sum_{i=1}^{N}\frac{\delta_i}{\sqrt{1-\delta_i}}}{\sqrt{\delta'\left\{\sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}\right\}}}$$

It follows that $\alpha$ has a Cauchy distribution as long as $\sigma_1 \neq 1$, $\sigma_2 \neq 1$, or $\kappa \pm 1$ (see, e.g., Cedilnik et al. 2004). These conditions are very mild under the Gaussian model. For instance, if no forecaster knows as much as the oracle, the conditions are satisfied. Consequently, the probability density function of $\alpha$ is

$$f(\alpha|x_0,\gamma) = \frac{1}{\pi}\frac{\gamma}{(\alpha-x_0)^2+\gamma^2},$$

where $x_0 = \kappa\sigma_1/\sigma_2$ and $\gamma = \sqrt{1-\kappa^2}\sigma_1/\sigma_2$. The parameter $x_0$ represents the location (the median and mode) and $\gamma$ specifies the scale (half the interquartile range) of the Cauchy distribution. The location parameter simplifies to

$$x_0 = \kappa\frac{\sigma_1}{\sigma_2} = \frac{N\sum_{i=1}^{N}\frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}}}{\sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}}}$$

Given that all the remaining terms are positive, the location parameter $x_0$ is also positive. Compare the $N$ terms with a given subindex $i$ in the numerator with the corresponding terms in the denominator. From $\delta' \geq \delta_i \geq \rho_{ij}$, it follows that

$$\frac{\delta_i}{1-\delta_i} = \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \tag{A.3}$$

$$\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}} \leq \frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \tag{A.4}$$

Therefore

$$N\sum_{i=1}^{N}\frac{\delta_i}{\sqrt{(1-\delta_i)(1-\delta')}} \geq \sum_{i=1}^{N}\frac{\delta_i}{1-\delta_i} + 2\sum_{i,j:i<j}\frac{\rho_{ij}}{\sqrt{(1-\delta_j)(1-\delta_i)}},$$

which gives that $x_0 \geq 1$. Given that the Cauchy distribution is symmetric around $x_0$, it must be the case that $\mathbb{P}(\alpha > 1|\boldsymbol{\Sigma}_{22}, \delta') \geq 1/2$. Based on (A.3) and (A.4), the location $x_0 = 1$ only when all the forecasters know the same information, i.e., when $\delta_i = \delta_j$ for all $i \neq j$. Under this particular setting, the amount of extremizing $\alpha$ is non-random and always equal to one. Any deviation from this particular information structure makes $\alpha$ random, $x_0 > 1$, and hence $\mathbb{P}(\alpha > 1|\boldsymbol{\Sigma}_{22}, \delta') > 1/2$. □

### A.3.4 Derivation of Equation 4.4

Clearly, any $\delta \in [0, 1]$ is plausible. Conditional on such $\delta$, however, the overlap parameter $\lambda$ must be within a subinterval of $[0, 1]$. The upper bound of this subinterval is always one because the forecasters may use the same information under any $\delta$ and $N$. To derive the lower bound, note that information overlap is unavoidable when $\delta > 1/N$, and that minimum overlap occurs when all information is used either by everyone or by a single forecaster. In other words, if $\delta > 1/N$ and $B_i \cap B_j = B$ with $|B| = \lambda\delta$ for all $i \neq j$, the value of $\lambda$ is minimized when $\lambda\delta + N(\delta - \delta\lambda) = 1$. Therefore the lower bound

for $\lambda$ is $\max\{(N-\delta^{-1})/(N-1), 0\}$, and $\mathbf{\Sigma}_{22}$ is coherent if and only if $\delta \in [0,1]$ and $\lambda|\delta \in [\max\{(N-\delta^{-1})/(N-1), 0\}, 1]$.

### A.3.5 Proof of Proposition 4.5.1

(i) This follows from direct computation:

$$\begin{aligned}\alpha &= \left(\frac{\frac{1}{(N-1)\lambda+1}\sum_{i=1}^{N} X_{B_i}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}}\right) \Bigg/ \left(\frac{1}{N}\sum_{i=1}^{N}\frac{X_{B_i}}{\sqrt{1-\delta}}\right) \\ &= \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}}, \end{aligned} \tag{A.5}$$

which simplifies to the given expression after substituting in $\gamma$. Given that this quantity does not depend on any $X_{B_i}$, it is non-random.

(ii) For a given $\delta$, the amount of extremizing $\alpha$ is minimized when $(N-1)\lambda+1$ is maximized. This happens as $\lambda \uparrow 1$. Plugging this into (A.5) gives

$$\alpha = \frac{\frac{N\sqrt{1-\delta}}{(N-1)\lambda+1}}{\sqrt{1-\frac{N\delta}{(N-1)\lambda+1}}} \downarrow \frac{\sqrt{1-\delta}}{\sqrt{1-\delta}} = 1$$

(iii) Assume without loss of generality that $\bar{P} > 0$. If $\max\{p_1, p_2, \ldots, p_N\} < 1$, then setting $\delta = 1/N$ and $\lambda = 0$ gives an aggregate probability $p'' = 1$ that is outside the convex hull of the individual probabilities. □

# Appendix B: Parameter Estimation Under Symmetric Information

This section describes how the maximum likelihood estimates of $\delta$ and $\lambda$ can be found accurately and efficiently. Denote a $N \times N$ matrix of ones with $\mathbf{J}_N$. A matrix $\mathbf{\Sigma}$ is called compound symmetric if it can be expressed in the form $\mathbf{\Sigma} = \mathbf{I}_N A + \mathbf{J}_N B$ for some constants $A$ and $B$. The inverse matrix (if it exists) and any scalar multiple of a compound symmetric matrix $\mathbf{\Sigma}$ are also compound symmetric (Dobbin and Simon, 2005). More specifically, for some constant $c$,

$$\begin{aligned} c\mathbf{\Sigma} &= \mathbf{I}_N(cA) + \mathbf{J}_N(cB) \\ \mathbf{\Sigma}^{-1} &= \mathbf{I}_N \frac{1}{A} - \mathbf{J}_N \frac{B}{A(A+NB)} \end{aligned} \tag{A.6}$$

Define

$$\begin{aligned} \mathbf{\Sigma}_{22} &:= \mathrm{Cov}\,(\boldsymbol{X}) = \mathbf{I}_N A_X + \mathbf{J}_N B_X \\ \mathbf{\Sigma}_P &:= \mathrm{Cov}\,(\boldsymbol{P}) = \mathbf{\Sigma}_{22}/(1-\delta) = \mathbf{I}_N A_P + \mathbf{J}_N B_P \\ \mathbf{\Omega} &:= \mathbf{\Sigma}_P^{-1} = \mathbf{I}_N A_\Omega + \mathbf{J}_N B_\Omega \end{aligned} \tag{A.7}$$

To set up the optimization problem, observe that the Jacobian for the map $\boldsymbol{P} \to \Phi\,(\boldsymbol{P}) = (\Phi(P_1), \Phi(P_2), \ldots, \Phi(P_N))'$ is $J(\boldsymbol{P}) = (2\pi)^{-N/2} \exp\,(-\boldsymbol{P}'\boldsymbol{P}/2)$. If $h(\boldsymbol{P})$ denotes the multivariate Gaussian density of $\boldsymbol{P} \sim \mathcal{N}_N\,(\mathbf{0}, \mathbf{\Sigma}_P)$, the density for $\boldsymbol{p} = (p_1, p_2, \ldots, p_N)'$ is

$$f\,(\boldsymbol{p}|\delta, \lambda) \;\; = \;\; h(\boldsymbol{P})J(\boldsymbol{P})^{-1} \propto |\mathbf{\Sigma}_P|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{P}'\mathbf{\Sigma}_P^{-1}\boldsymbol{P}\right],$$

where $\boldsymbol{P} = \Phi^{-1}(\boldsymbol{p})$. Let $\mathbf{S}_P = \boldsymbol{P}\boldsymbol{P}'$ be the (rank one) sample covariance matrix of $\boldsymbol{P}$. The log-likelihood then reduces to

$$\log f\left(\boldsymbol{p}|\delta,\lambda\right) \propto -\log\det\boldsymbol{\Sigma}_P - \operatorname{tr}\left(\mathbf{S}_P^{-1}\boldsymbol{\Sigma}_P\right)$$

This log-likelihood is not concave in $\boldsymbol{\Sigma}_P$. It is, however, a concave function of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_P^{-1}$. Making this change of variables gives us the following optimization problem:

$$\begin{aligned} &\text{minimize } -\log\det\boldsymbol{\Omega} + \operatorname{tr}\left(\mathbf{S}_P\boldsymbol{\Omega}\right) \\ &\text{subject to } \delta \in [0,1] \\ &\qquad\qquad \lambda \in \left[\max\left\{\frac{N-\delta^{-1}}{N-1}, 0\right\}, 1\right), \end{aligned} \tag{A.8}$$

where the open upper bound on $\lambda$ ensures a non-singular information structure $\boldsymbol{\Sigma}_{22}$. Unfortunately, the feasible region is not convex (see, e.g., Figure 3 in the main manuscript) but can be made convex by re-expressing the problem as follows: First, let $\rho = \delta\lambda$ denote the amount of information known by a forecaster; that is, let $A_X = (\delta - \rho)$ and $B_X = \rho$. Solving the problem in terms of $\delta$ and $\rho$ is equivalent to minimizing the original objective (A.8) but subject to $0 \leq \rho \leq \delta$ and $0 \leq \rho(N-1) - N\delta + 1$. Given that this region is an intersection of four half-spaces, it is convex. Furthermore, it can be translated into the corresponding feasible and convex set of $(A_\Omega, B_\Omega)$ via the following steps:

$$\begin{aligned} & \boldsymbol{\Sigma}_{22} \in \{\boldsymbol{\Sigma}_{22} : 0 \leq \rho \leq \delta, 0 \leq \rho(N-1) - N\delta + 1\} \\ \Leftrightarrow \quad & \boldsymbol{\Sigma}_{22} \in \{\boldsymbol{\Sigma}_{22} : 0 \leq B_X, 0 \leq A_X, 0 \leq 1 - B_X + NA_X, \} \\ \Leftrightarrow \quad & \boldsymbol{\Sigma}_P \in \{\boldsymbol{\Sigma}_P : 0 \leq A_P \leq 1/(N-1), 0 \leq B_P\} \\ \Leftrightarrow \quad & \boldsymbol{\Omega} \in \{\boldsymbol{\Omega} : 0 \leq A_\Omega - N + 1, 0 \leq A_\Omega + B_\Omega N, 0 \leq -B_\Omega\} \end{aligned}$$

According to Rao (2009), $\log\det(\mathbf{\Omega}) = N\log A_\Omega + \log(1 + NB_\Omega/A_\Omega)$. Plugging this and the feasible region of $(A_\Omega, B_\Omega)$ into the original problem (A.8) gives an equivalent but convex optimization problem:

$$\begin{aligned}
&\text{minimize } -N\log A_\Omega - \log\left(1 + \frac{NB_\Omega}{A_\Omega}\right) + A_\Omega\,\mathrm{tr}(\mathbf{S}_P) + B_\Omega\,\mathrm{tr}(\mathbf{S}_P\mathbf{J}_N)\\
&\text{subject to } 0 \le A_\Omega - N + 1\\
&\qquad\qquad 0 \le A_\Omega + B_\Omega N\\
&\qquad\qquad 0 \le -B_\Omega
\end{aligned}$$

The first term of this objective is both convex and non-decreasing. The second term is a composition of the same convex, non-decreasing function with a function that is concave over the feasible region. Such a composition is always convex. The last two terms are affine and hence also convex. Therefore, given that the objective is a sum of four convex functions, it is convex, and globally optimal values of $(A_\Omega, B_\Omega)$ can be found very efficiently with interior point algorithms such as the barrier method. There are many open software packages that implement generic versions of these methods. For instance, our implementation uses the standard R function `constrOptim` to solve the optimization problem. Denote optimal values with $(A^*_\Omega, B^*_\Omega)$. They can be traced back to $(\delta, \lambda)$ via (A.6) and (A.7). The final map simplifies to

$$\delta^* = \frac{B^*_\Omega(N-1) + A^*_\Omega}{A^*_\Omega(1 + A^*_\Omega) + B^*_\Omega(N - 1 + NA^*_\Omega)} \qquad \text{and} \qquad \lambda^* = -\frac{B^*_\Omega}{B^*_\Omega(N-1) + A^*_\Omega}$$

## A.4 Supplement for Chapter 5

# Appendix A: Proofs and Derivations

### A.4.1 Proof of Proposition 5.2.1

*Proof.* This follows from direct computation and the definition of conditional expectation as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)\mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\middle|\mathcal{F}_j\right) &= \mathbb{E}_{\mathbb{Q}}\left(\mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)\frac{d\mathbb{P}}{d\mathbb{Q}}\middle|\mathcal{F}_j\right) \quad (\text{since } \mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j) \in \mathcal{F}_j) \\
\Leftrightarrow \int_A \mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)\mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\middle|\mathcal{F}_j\right) d\mathbb{Q} &= \int_A \mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)\frac{d\mathbb{P}}{d\mathbb{Q}}d\mathbb{Q} \qquad (\text{for all } A \in \mathcal{F}_j) \\
&= \int_A \mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)d\mathbb{P} \\
&= \int_A Y d\mathbb{P} \\
&= \int_A \frac{d\mathbb{P}}{d\mathbb{Q}} Y d\mathbb{Q},
\end{aligned}$$

which then gives that $\mathbb{E}_{\mathbb{Q}}(\frac{d\mathbb{P}}{d\mathbb{Q}}Y|\mathcal{F}_j) = \mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j)\mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}|\mathcal{F}_j\right)$. Dividing both sides by $\mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}|\mathcal{F}_j\right)$ gives the final answer. □

### A.4.2 Proof of Proposition 5.2.2

*Proof.* Denote the mean of the target outcome $Y$ with $\mu_0 := \mathbb{E}(Y)$. Each item is then proved as follows.

i) Given that $\mathbb{E}(Y|X_j) = X_j$, the law of iterated expectation gives

$$\mathbb{E}(X_j) = \mathbb{E}(\mathbb{E}(Y|X_j)) = \mathbb{E}(Y)$$

for all $j = 1, \ldots, N$.

iii)

$$\begin{aligned}
\mathrm{Cov}\,(X_j, X_i) &= \mathbb{E}((X_j - \mu_0)(X_i - \mu_0)) \\
&= \mathbb{E}(X_j X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(X_j|X_i)X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(\mathbb{E}(Y|X_j)|X_i)X_i) - \mu_0^2 \\
&= \mathbb{E}(\mathbb{E}(Y|X_i)X_i) - \mu_0^2 && \text{(the smallest } \sigma\text{-field wins)} \\
&= \mathbb{E}(X_i^2) - \mu_0^2 \\
&= \mathrm{Var}\,(X_i)
\end{aligned}$$

ii)

$$\begin{aligned}
\mathrm{Var}\,(X_i) &= \mathrm{Cov}\,(X_i, X_j) && \text{(by item iii)} \\
&= \mathbb{E}((X_i - \mu_0)(X_j - \mu_0)) \\
&\leq \mathbb{E}((X_i - \mu_0)^2)^{1/2}\mathbb{E}((X_j - \mu_0)^2)^{1/2} && \text{(by Cauchy-Schwarz' inequality)} \\
&= \sqrt{\mathrm{Var}\,(X_i)\mathrm{Var}\,(X_j)},
\end{aligned}$$

which then provides $\mathrm{Var}\,(X_i) \leq \mathrm{Var}\,(X_j)$. This inequality is tight because $X_i = X_j$ for $\mathcal{F}_i = \mathcal{F}_j$.

□

### A.4.3 Finding $\mu^*$ for $\mathcal{P}_{sd}(\cdot : \kappa)$

This section describers a binary-search-like algorithm to solve

$$\mu^* = \arg\min_{\mu \geq 0} \pi(\mu) = \arg\min_{\mu \geq 0} \sum_{i=1}^{N} \left((\mu - l_i)_+^2 + (l_i - \kappa\mu)_+^2\right) \tag{A.9}$$

**Algorithm 3** This procedure solves (A.9) efficiently using the structure of the problem and binary-search.

**Require:** Condition number threshold $\kappa \geq 1$ and sample eigenvalues in ascending order $l_1 \leq l_2 \leq \cdots \leq l_{N+1}$.

**procedure** BINARY-SEARCH OPTIMIZATION
  Initialize $D \leftarrow \max\{l_1, 0\}$ and $U \leftarrow l_{N+1}/\kappa$.
  $\mu_0 \leftarrow (D+U)/2$
  **for** $n = 0, 1, \ldots$ **do**
    Compute $\mu_n^*$, $\mathfrak{d}_n$, and $\mathfrak{u}_n$.
    **if** $\mu_n^* < 0$ and $\mathfrak{d}_n < 0$ **then**
      **return** $0$
    **else if** $\mu_n^* < \mathfrak{d}_n$ **then**
      $U \leftarrow \mathfrak{d}_n$
    **else if** $\mu_n^* > \mathfrak{u}_n$ **then**
      $D \leftarrow \mathfrak{u}_n$
    **else**
      **return** $\mu_n^*$
    **end if**
    $\mu_{n+1} \leftarrow (D+U)/2$
  **end for**
  **return** $\mu_n^*$
**end procedure**

First, it can be assumed that $\text{cond}(h(\mathbf{S}_Z)) \notin [1, \kappa]$; otherwise, the projection can simply return $h(\mathbf{S}_Z)$. Second, $\max\{0, l_1\} \leq \mu \leq l_{N+1}/\kappa$ because otherwise moving $\mu$ closer to the nearest sample eigenvalue decreases $\pi(\mu)$. Now, consider some value $\mu_n \geq 0$ and two index sets $\mathfrak{D}_n = \{i : l_i \leq \mu_n\}$ and $\mathfrak{U}_n = \{i : \mu_n \kappa \leq l_i\}$. Then,

$$\pi(\mu_n) = \sum_{i \in \mathfrak{D}_n} (\mu_n - l_i)^2 + \sum_{i \in \mathfrak{U}_n} (l_i - \kappa \mu_n)^2 ,$$

which has a global minimum at

$$\mu_n^* = \frac{\sum_{i \in \mathfrak{D}_n} l_i + \kappa \sum_{i \in \mathfrak{U}_n} l_i}{|\mathfrak{D}_n| + \kappa^2 |\mathfrak{U}_n|}$$

The operator $|\mathfrak{A}|$ denotes the number of elements in the set $\mathfrak{A}$. Let $\mathfrak{d}_n$ and $\mathfrak{u}_n$ denote the minimum and maximum, respectively, of the interval where any value of $\mu$ gives the index

sets $\mathfrak{D}_n$ and $\mathfrak{U}_n$. To make this specific, define two operators:

$$d(\mu) = \max\{l_i : l_i \leq \mu\} \qquad \text{and} \qquad u(\mu) = \min\{l_i : l_i \geq \mu\}.$$

If no value is found, then $d(\mu) = 0$ and $u(\mu) = +\infty$. Then,

$$\mathfrak{d}_n = \max\{d(\mu_n), d(\mu_n\kappa)/\kappa\}$$
$$\mathfrak{u}_n = \min\{u(\mu_n), u(\mu_n\kappa)/\kappa\}$$

Of course, $\mu_n^*$ is the solution to (A.9) as long as $\mu_n^* \in (\mathfrak{d}_n, \mathfrak{u}_n]$. If, on the other hand, $\mu_n^*$ is less than $\mathfrak{d}_n$ (or greater than $\mathfrak{u}_n$), the global minimum $\mu^*$ must be smaller than $\mathfrak{d}_n$ (or greater than $\mathfrak{u}_n$). If $\mu_n^*$ is, say, less than $\mathfrak{d}_n$, then a natural approach is to update $\mu_n$ to $\mu_{n+1}$ that is somewhere between $\mathfrak{d}_n$ and some known lower bound of $\mu$. This gives rise to a binary-search-like algorithm described in Algorithm 3.

# Appendix B: Synthetic Data Analysis

This supplementary section evaluates the models under synthetic data generated directly from the multivariate Gaussian distribution (7.3). The analysis provides insight into the behavior of the estimation procedure and also introduces the simplest instance of the Gaussian model.

> **Model Instance.** The link function $g(\cdot)$ is the identity. Thus, the target quantity is $Y_k = g(Z_{0k}) = Z_{0k}$, and the forecasts are $X_{jk} = \mathbb{E}(Y_k|Z_{jk}) = Z_{jk}$ for all $j$ and $k$. The revealed aggregator for event $k$ is $X_k'' = \text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}\mathbf{X}_k$, where $\mathbf{X}_k = (X_{1k}, \ldots, X_{Nk})'$.

Simulating forecasts from (7.3) requires a $\mathbf{\Sigma}$ such that $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$. One approach is to generate a random $N \times N$ positive definite matrix from a Wishart distribution, scale it

such that all diagonal entries are within $[0, 1]$, and accept this scaled version if it satisfies $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$. However, based on a brief simulation study that is not presented here for the sake of brevity, the rate at which the randomly generated matrix is accepted decreases in $N$ and is very close to zero already for $N > 5$. Therefore this section adopts a different approach that samples $\mathbf{\Sigma}$ with full acceptance rate but only within a subset of all information structures: first pick $\delta_j \overset{i.i.d.}{\sim} \mathcal{U}(0.1, 0.9)$ and then set $\rho_{ij} = \delta_i \delta_j$ for all $i \neq j$. This way $\mathbf{\Sigma} - \text{diag}(\mathbf{\Sigma})\text{diag}(\mathbf{\Sigma})' = \text{Diag}((\delta_1 - \delta_1^2, \ldots, \delta_N - \delta_N^2)') \in \mathcal{S}_+^N$, which, by the Schur complement, satisfies $h(\mathbf{\Sigma}) \in \mathcal{S}_+^{N+1}$. Finally, the outcome and forecasts for the $k$th event are drawn from $(Y_k, \mathbf{X}_k) = (Z_{0k}, Z_{1k}, \ldots, Z_{Nk})' \overset{i.i.d.}{\sim} \mathcal{N}_{N+1}(\mathbf{0}, h(\mathbf{\Sigma}))$. These forecasts are aggregated in the following ways:

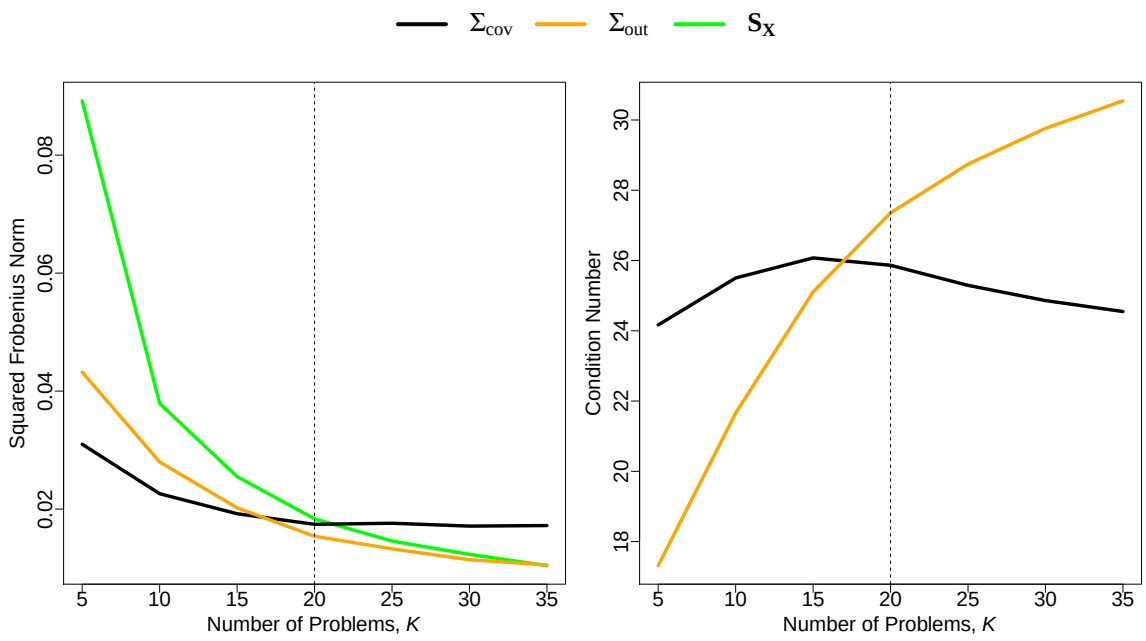
1. $X_k''(\mathbf{S_X}) = \text{diag}(\mathbf{S_X})'\mathbf{S_X}^{-1}\mathbf{X}_k$, where $\mathbf{S_X}$ is the sample covariance matrix. Given that $\mathbf{S_X}$ is singular when $K < N$, its inverse is computed with the (Moore-Penrose) generalized inverse.

2. $X_{cov}'' = X_k''(\mathbf{\Sigma}_{cov}) = \text{diag}(\mathbf{\Sigma}_{cov})'\mathbf{\Sigma}_{cov}^{-1}\mathbf{X}_k$, where $\mathbf{\Sigma}_{cov} = \mathcal{P}_{LSE}(\mathbf{S_X} : \kappa_{cov})$. The condition number constraint $\kappa_{cov}$ is found over a grid of $100$ values between $10$ and $1,000$.

3. $X_{out}'' = X_k''(\mathbf{\Sigma}_{out}) = \text{diag}(\mathbf{\Sigma}_{out})'\mathbf{\Sigma}_{out}^{-1}\mathbf{X}_k$, where $\mathbf{\Sigma}_{out} = \mathcal{P}_{LSE}(\mathbf{S_X} : \kappa_{out})$ and $\kappa_{out}$ is found using cross-validation as proposed by Won and Kim (2006). More specifically, they suggest choosing $\kappa$ by maximizing the expected predictive log-likelihood. More specifically, if $\mathbf{Z}_1, \ldots, \mathbf{Z}_K, \tilde{\mathbf{Z}} \overset{i.i.d.}{\sim} \mathcal{N}_N(\mathbf{0}, \mathbf{\Sigma})$, they recommend using

$$\kappa = \underset{\nu \geq 1}{\arg\max}\, \mathbb{E}\left\{\mathbb{E}_{\tilde{\mathbf{Z}}}\left[\ell(\tilde{\mathbf{Z}}, \mathcal{P}_{LSE}(\mathbf{S} : \nu))\right]\right\},$$

where $\ell(\cdot)$ denotes the log-likelihood and $\mathbf{S}$ is computed only based on $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$. They approximate the expected predictive log-likelihood with cross validation. This partitions the data $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_K)'$ into $R$ subsets such that $\mathbf{Z} = \left(\mathbf{Z}_{(1)}, \ldots, \mathbf{Z}_{(R)}\right)'$.




(a) Estimation accuracy of $\mathbf{\Sigma}$

(b) Average condition numbers

Figure A.2: Estimation of the information structure and the average condition numbers of the estimates. Both are important for accurate prediction of $Y_k$. The vertical dashed lines represents the number of forecasters fixed at $N = 20$.

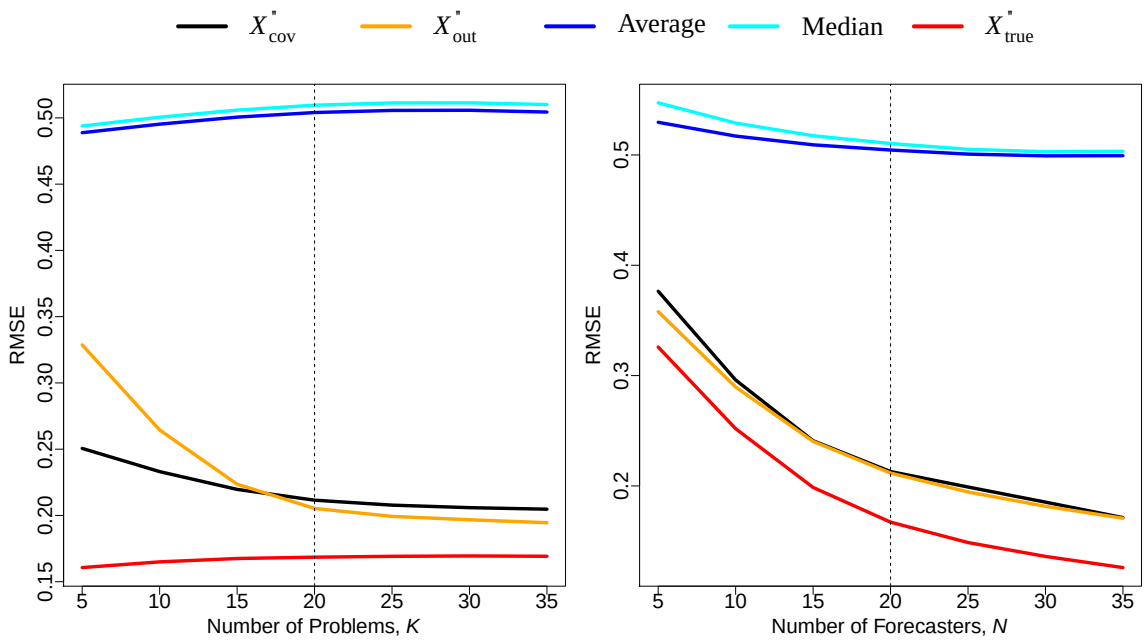
During the $r$th iteration, $\mathbf{Z}_{(r)}$ functions as $\tilde{\mathbf{Z}}$ and the remaining $R-1$ subsets form the estimated covariance matrix, denoted with $\mathbf{S}_{(r)}$. In this section, the cross-validation uses five folds and $\kappa_{out}$ is found using the same grid as in $X''_{cov}$.

4. $X''_{true} = X''_k(\mathbf{\Sigma}) = \text{diag}(\mathbf{\Sigma})'\mathbf{\Sigma}^{-1}\mathbf{X}_k$. This aggregator assumes the knowledge of the true $\mathbf{\Sigma}$ and hence represents optimal performance.
5. The average forecast
6. The median forecast

The overall process is repeated $5,000$ times under different values of $K$ and $N$, each ranging from $5$ to $35$ with constant increments of $5$. The final results then represent average performance across those $5,000$ iterations.

Recall that accurate revealed aggregation arises from a precise estimate of $\mathbf{\Sigma}$ and a low condition number. This allows different strategies for achieving good aggregation. In fact, the two selection procedures discussed in Section 5.3.4 and item 3 above make slightly different tradeoffs. This is illustrated in Figure A.2 that varies $K$ between $5$ and $35$ but keeps $N$ fixed at $20$. More specifically, Figure A.2a examines how $\mathbf{\Sigma}_{cov}$, $\mathbf{\Sigma}_{out}$, and $\mathbf{S_X}$ capture $\mathbf{\Sigma}$. Even though all estimators become more accurate as $K$ grows, $\mathbf{\Sigma}_{out}$ and $\mathbf{S_X}$ improve at a higher rate than $\mathbf{\Sigma}_{cov}$. In fact, if $K > N$, $\mathbf{S_X}$ and $\mathbf{\Sigma}_{out}$ perform better than $\mathbf{\Sigma}_{cov}$. On the other hand, if $K < N$, $\mathbf{\Sigma}_{cov}$ is more accurate than the other two. Figure A.2b presents the corresponding (average) condition numbers of these estimates. This plot omits cond$(\mathbf{S}_Z)$ because this was overall very large and hence made the scale too wide for a proper comparison of cond$(\mathbf{\Sigma}_{cov})$ and cond$(\mathbf{\Sigma}_{out})$. Notice that in this figure cond$(\mathbf{\Sigma}_{out})$ increases while cond$(\mathbf{\Sigma}_{cov})$ generally decreases as $K$ grows larger. In fact, when $K > N$, cond$(\mathbf{\Sigma}_{cov})$ is smaller than cond$(\mathbf{\Sigma}_{out})$. From the prediction perspective, this makes conditional-validation more forgiving towards error in the estimated $\mathbf{\Sigma}$. Therefore, while $\mathbf{\Sigma}_{cov}$ incorporates the actual prediction process and looks for a fine balance between a precise estimate of $\mathbf{\Sigma}$ and a low condition number, $\mathbf{\Sigma}_{out}$ is unaware of the details of revealed aggregation and hence simply focuses on estimating $\mathbf{\Sigma}$ as accurately as possible.

These two strategies lead to slightly different predictive behavior as is illustrated in Figure A.3. This plot shows the average RMSEs of the competing aggregators in predicting $Y_k$. Figure A.3a varies $K$ but fixes $N = 20$. Figure A.3b, on the other hand, varies $N$ but fixes $K = 20$. Given that $Y_k = Z_{0k} \sim \mathcal{N}(0, 1)$, the RMSE of the prior mean $\mathbb{E}(\sqrt{(Y_k - 0)^2}) = \mathbb{E}(|Y_k|) = \sqrt{2/\pi} \approx 0.8$ can be considered as the upper bound in prediction error. The lower bound, on the other hand, is given by $X''_{true}$. The revealed aggregator $X''(\mathbf{S_X})$ typically received a loss much larger than $0.8$ and is therefore not included in the figure. Overall, the two measurement-error aggregators, namely average and median perform very similarly, with RMSE around $0.5$. They both show slight improve-




(a) Prediction accuracy under fixed $N = 20$ but different values of $K$.

(b) Prediction accuracy under fixed $K = 20$ but different values of $N$.

Figure A.3: The accuracy to predict $Y_k$ under different values of $N$ and $K$. The aggregator $X''_{true}$ assumes knowledge of the true information structure and hence represents optimal accuracy.

ment as $N$ increases. In all cases, however, their RMSE is uniformly well above that of the revealed aggregators, suggesting that measurement-error aggregators are a poor choice when forecasts truly arise from a partial information model. The revealed aggregators $X''_{cov}$ and $X''_{out}$ perform very similarly when $K \geq 15$. They collect information and appear to improve at the optimal rate as $N$ increases. This can be seen in the way the performance gap from $X_{true}$ to $X''_{out}$ and $X''_{cov}$ remains approximately constant in Figure A.3b. They both, however, approach $X''_{true}$ as $K$ grows larger. When $K$ is small, say less than 15, $X''_{cov}$ is more robust and clearly yields better results than $X''_{out}$. This is an important consideration because in prediction polling $K$ is often much smaller than $N$. For all these reasons only $X''_{cov}$ is considered in this paper.

## A.5 Supplement for Chapter 6

### A.5.1 Proof of Proposition 6.2.1

Under the Gaussian model the joint distribution of $X_S, X_{B_1}$ and $X_{B_2}$ is

$$\begin{pmatrix} X_S \\ X_{B_1} \\ X_{B_2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$

where

$$\begin{pmatrix} \Sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} := \left(\begin{array}{c|cc} 2 & 1 & 1 \\ \hline 1 & 1 & \rho \\ 1 & \rho & 1 \end{array}\right).$$

The inverse of $\boldsymbol{\Sigma}_{22}$ is

$$\boldsymbol{\Sigma}_{22}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Using the well-known properties of a conditional multivariate Gaussian distribution (see, e.g., Ravishanker and Dey 2001, Result 5.2.10), the distribution of $X_S$ given $\boldsymbol{X} = (X_{B_1}, X_{B_2})'$ is $X_S|\boldsymbol{X} \sim \mathcal{N}(\mu_S, \sigma_S^2)$, where

$$\begin{aligned} \mu_S &= \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X} = \frac{1}{1+\rho}(X_{B_1} + X_{B_2}), \text{ and} \\ \sigma_S^2 &= \Sigma_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \frac{2\rho}{1+\rho}. \end{aligned}$$

Denoting $p^{(1)}$ and $p^{(2)}$ respectively by $p$ and $q$, we recall that the individual forecasts are $p = \Phi(X_{B_1})$ and $q = \Phi(X_{B_2})$. The synthesized forecast is then

$$\begin{aligned} g_\rho(p,q) &= \mathbb{P}(X_S > 0 | p, q) \\ &= \mathbb{P}(X_S > 0 | X_{B_1}, X_{B_2}) \\ &= 1 - \Phi\left( \frac{-\frac{1}{1+\rho}(X_{B_1} + X_{B_2})}{\sqrt{\frac{2\rho}{1+\rho}}} \right) \\ &= \Phi\left( \frac{\Phi^{-1}(p) + \Phi^{-1}(q)}{\sqrt{2\rho(1+\rho)}} \right). \end{aligned}$$

### A.5.2 Proof of Proposition 6.2.1

To compute $\lambda_\rho(p,q)$, recall that $Z_1$ and $Z_2$ are standard normals with covariance $\rho$ and that $(Z_1, Z_2)$ maps to $(p,q)$ by $\Phi$ in each coordinate. The density of $(Z_1, Z_2)$ at $(x,y)$ is proportional to

$$(2\pi)^{-1}(\det Q)^{1/2} \exp\left[\frac{1}{2} Q(x,y)\right]$$

where the quadratic form $Q$ is the inverse of the covariance matrix:

$$Q = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

Thus the density $h(x,y)$ of $(Z_1, Z_2)$ at $(x,y)$ is equal to

$$\frac{1}{2\pi}(1-\rho^2)^{-1/2} \exp\left[-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right]. \tag{A.10}$$

The Jacobian of the map $(x,y) \mapsto (\Phi(x), \Phi(y))$ at $(x,y)$ is given by

$$\frac{1}{2\pi} \exp\left[-\frac{1}{2}\left(x^2 + y^2\right)\right] \tag{A.11}$$

and therefore

$$\begin{aligned}\lambda_\rho(p,q) &= h(x,y)J(x,y)^{-1}\big|_{x=\Phi^{-1}(p),y=\Phi^{-1}(q)} \\ &= c(1-\rho^2)^{-1/2}\exp\left[-\frac{\rho^2x^2-2\rho xy+\rho^2y^2}{2(1-\rho^2)}\right].\end{aligned}$$

Putting this together with (6.3) and (6.4) gives

$$g(p,q) = A_0/B_0, \tag{A.12}$$

where

$$\begin{aligned}A_0 &= \int \Phi\left(\frac{\Phi^{-1}(p)+\Phi^{-1}(q)}{\sqrt{2\rho(1+\rho)}}\right)(1-\rho^2)^{-1/2} \\ &\quad\times\exp\left[-\frac{\rho^2\Phi^{-1}(p)^2-2\rho\Phi^{-1}(p)\Phi^{-1}(q)+\rho^2\Phi^{-1}(q)^2}{2(1-\rho^2)}\right]d\rho \\ B_0 &= (1-\rho^2)^{-1/2}\exp\left[-\frac{\rho^2\Phi^{-1}(p)^2-2\rho\Phi^{-1}(p)\Phi^{-1}(q)+\rho^2\Phi^{-1}(q)^2}{2(1-\rho^2)}\right]\end{aligned}$$

By symmetry, we may assume without loss of generality that $p < q$. Removing a factor of $\exp\left[\frac{1}{2}\left(\Phi^{-1}(p)^2+\Phi^{-1}(q)^2\right)\right]$ from both numerator and denominator of (A.12) gives

$$g(p,q) = A_1/B_1, \tag{A.13}$$

where

$$\begin{aligned}A_1 &= \int_0^1 \Phi\left(\frac{\Phi^{-1}(p)+\Phi^{-1}(q)}{\sqrt{2\rho(1+\rho)}}\right)\frac{1}{\sqrt{1-\rho^2}} \\ &\quad\exp\left(-\frac{\Phi^{-1}(p)^2-2\rho\Phi^{-1}(p)\Phi^{-1}(q)+\Phi^{-1}(q)^2}{2(1-\rho^2)}\right)d\rho \\ B_1 &= \int_0^1 \frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{\Phi^{-1}(p)^2-2\rho\Phi^{-1}(p)\Phi^{-1}(q)+\Phi^{-1}(q)^2}{2(1-\rho^2)}\right)d\rho\end{aligned}$$

We compute first the denominator of (A.13), then the numerator, which uses similar techniques but is a little more involved.

**Computation of the denominator**

Denote the denominator of (A.13) by

$$I_2 := \int_0^1 \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\Phi^{-1}(p)^2 - 2\rho\Phi^{-1}(p)\Phi^{-1}(q) + \Phi^{-1}(q)^2}{2(1-\rho^2)}\right) d\rho\,. \qquad \text{(A.14)}$$

Denote the density, CDF and tail of the bivariate standard normal with correlation parameter $\rho \in (-1, 1)$ respectively by

$$\begin{aligned}
\phi_2(x, y; \rho) &= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}} \\
\Phi_2(b_1, b_2; \rho) &= \int_{-\infty}^{b_1}\int_{-\infty}^{b_2} \phi_2(x, y; \rho) dy dx \\
L(b_1, b_2, \rho) &= \Phi_2(-b_1, -b_2, \rho)\,.
\end{aligned}$$

Plackett's formula (Plackett, 1954) gives

$$\frac{\partial L(b_1, b_2, \rho)}{\partial \rho} = \frac{\exp\left(-\frac{b_1^2 - 2\rho b_1 b_2 + b_2^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}}$$

specializes to the integrand in (A.14) when $b_1 = \Phi^{-1}(p)$ and $b_2 = \Phi^{-1}(q)$, whence

$$I_2 = \int_0^1 2\pi \frac{\partial}{\partial \rho} L\left(\Phi^{-1}(p), \Phi^{-1}(q), \rho\right) d\rho\,.$$

Using the identities $L(b_1, b_2, 0) = \Phi(-b_1)\Phi(-b_2)$ and $L(b_1, b_2, 1) = \Phi(-\max\{b_1, b_2\})$ along with $p < q$ gives

$$\begin{aligned} I_2 &= 2\pi \left[ L(\Phi^{-1}(p), \Phi^{-1}(q), 1) - L(\Phi^{-1}(p), \Phi^{-1}(q), 0) \right] \\ &= 2\pi \left[ \Phi(-\max\{\Phi^{-1}(p), \Phi^{-1}(q)\}) - \Phi(-\Phi^{-1}(p))\Phi(-\Phi^{-1}(q)) \right] \\ &= 2\pi(1-q)p\,. \end{aligned} \tag{A.15}$$

**Computation of the numerator**

Denote the numerator of (A.13)

$$I_1 = \int_0^1 \Phi\left( \frac{\Phi^{-1}(p) + \Phi^{-1}(q)}{\sqrt{2\rho(1+\rho)}} \right) \frac{1}{\sqrt{1-\rho^2}} \tag{A.16}$$

$$\times \exp\left( -\frac{\Phi^{-1}(p)^2 - 2\rho\Phi^{-1}(p)\Phi^{-1}(q) + \Phi^{-1}(q)^2}{2(1-\rho^2)} \right) d\rho\,. \tag{A.17}$$

Extending the notation from before, denote the trivariate normal CDF by

$$\Phi_3(b_1, b_2, b_3; R) = \frac{1}{(2\pi)^{3/2}|R|^{1/2}} \int_{-\infty}^{b_1} \int_{-\infty}^{b_2} \int_{-\infty}^{b_3} \exp\left( -\frac{x^T R^{-1} x}{2} \right) dx_3\, dx_2\, dx_1, \tag{A.18}$$

where $R = (\rho_{ij})$ is the correlation matrix. In Plackett (1954) there is a formula as well for the partial derivative of the trivariate CDF with respect to the coefficient $\rho_{12}$, meaning that the $(1,2)$ and $(2,1)$ entries of $R$ change while all other entries remain constant:

$$\frac{\partial \Phi_3(b_1, b_2, b_3; R)}{\partial \rho_{12}} = \frac{\exp\left( -\frac{b_1^2 - 2\rho_{12} b_1 b_2 + b_2^2}{2(1-\rho^2)} \right)}{2\pi\sqrt{1-\rho_{12}^2}} \Phi(u_3(\rho_{12}))\,, \tag{A.19}$$

where

$$u_3(\rho) = \frac{b_3(1-\rho^2) - b_1(\rho_{31} - \rho\rho_{32}) - b_2(\rho_{32} - \rho\rho_{31})}{\sqrt{(1-\rho^2)(1-\rho^2-\rho_{31}^2-\rho_{32}^2+2\rho\rho_{31}\rho_{32})}}\,. \tag{A.20}$$

Plugging in

$$b_1 = -\Phi^{-1}(p), \quad b_2 = -\Phi^{-1}(q), \quad b_3 = 0, \quad \text{and } \rho_{31} = \rho_{32} = \frac{1}{\sqrt{2}}$$

gives

$$u_3(\rho_{12}) = \frac{\frac{1-\rho_{12}}{\sqrt{2}}\left(\Phi^{-1}(p) + \Phi^{-1}(q)\right)}{\sqrt{(1-\rho_{12}^2)(\rho_{12}-\rho_{12}^2)}} = \frac{\Phi^{-1}(p) + \Phi^{-1}(q)}{\sqrt{2\rho_{12}(1+\rho_{12})}}$$

leading to

$$\frac{\partial \Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; \begin{pmatrix} 1 & \rho_{12} & \sqrt{1/2} \\ \rho_{12} & 1 & \sqrt{1/2} \\ \sqrt{1/2} & \sqrt{1/2} & 1 \end{pmatrix}\right)}{\partial \rho_{12}}$$

$$= \frac{\exp\left(-\frac{\Phi^{-1}(p)^2 - 2\rho_{12}\Phi^{-1}(p)\Phi^{-1}(q) + \Phi^{-1}(q)^2}{2(1-\rho_{12}^2)}\right)}{2\pi\sqrt{1-\rho_{12}^2}} \times \Phi\left(\frac{\Phi^{-1}(p) + \Phi^{-1}(q)}{\sqrt{2\rho_{12}(1+\rho_{12})}}\right). \tag{A.21}$$

Integrating (A.21) as $\rho_{12}$ goes from 0 to 1 and comparing to (A.16) we see that

$$I_1 \tag{A.22}$$

$$= 2\pi \int_0^1 \frac{\partial}{\partial \rho_{12}} \Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; \begin{pmatrix} 1 & \rho_{12} & \sqrt{1/2} \\ \rho_{12} & 1 & \sqrt{1/2} \\ \sqrt{1/2} & \sqrt{1/2} & 1 \end{pmatrix}\right) d\rho_{12}$$

$$= 2\pi \left[\Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R\right) - \Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R^*\right)\right], \tag{A.23}$$

where the matrices $R,\ R^*$ are given by

$$R = \begin{pmatrix} 1 & 1 & \frac{1}{\sqrt{2}} \\ 1 & 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{pmatrix}, \qquad R^* = \begin{pmatrix} 1 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{pmatrix}. \tag{A.24}$$

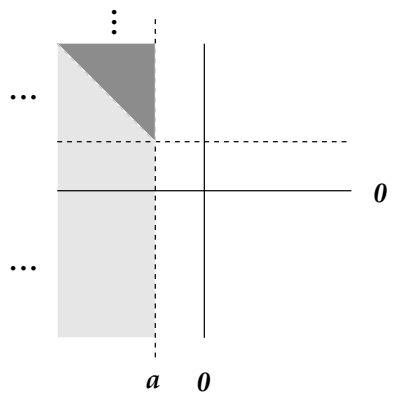



Figure A.4: The darker region has probability $\mathbb{P}(Y_1 \leq a)^2/2$

Computing first $\Phi_3(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R)$, we remark that $R$ forces $X_1 = X_2$, whence $\Phi_3(a, b, c; R) = \Phi_2(-\max\{a, b\}, c; R')$ where $R' = \begin{pmatrix} 1 & \sqrt{1/2} \\ \sqrt{1/2} & 1 \end{pmatrix}$. If $(X_1, X_2)$ is Gaussian with covariance $R'$ then $X_1 = Y_1$ and $X_2 = (Y_1 + Y_2)/\sqrt{2}$ where $(Y_1, Y_2)$ are independent standard normals. Thus, using $p < q$,

$$\begin{aligned} \Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R\right) &= \Phi_2\left(-\Phi^{-1}(q), 0; R'\right) \\ &= \mathbb{P}(X_1 \leq -\Phi^{-1}(q), X_2 \leq 0) \\ &= \mathbb{P}(Y_1 \leq -\Phi^{-1}(q), Y_2 \leq -Y_1)\,. \end{aligned}$$

Meyer (2009) remarks (see Figure A.4) that

$$\mathbb{P}(Y_1 \leq a, Y_2 \leq -Y_1) = \mathbb{P}(Y_1 \leq a) - \frac{1}{2}\mathbb{P}(Y_1 \leq a)^2\,.$$

Thus,

$$\Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R\right) = (1 - q) - \frac{(1 - q)^2}{2}\,. \tag{A.25}$$

Next, we compute $\Phi_3\left(-\Phi^{-1}(p), -\Phi^{-1}(q), 0; R^*\right)$. In this case,

$$(X_1, X_2, X_3) = (Y_1, Y_2, (Y_1 + Y_2)/\sqrt{2}),$$

where again $(Y_1, Y_2)$ is a pair if independent standard normals. We need therefore to compute

$$\mathbb{P}(Y_1 \leq -\Phi^{-1}(p), Y_2 \leq -\Phi^{-1}(q), Y_1 + Y_2 \leq 0).$$

We claim that

$$\begin{aligned} &\mathbb{P}(Y_1 \leq -\Phi^{-1}(p), Y_2 \leq -\Phi^{-1}(q), Y_1 + Y_2 \leq 0) \\ &= \begin{cases} (1-p)(1-q) & \text{if } p+q \geq 1; \\ \dfrac{1-p^2-q^2}{2} & \text{if } p+q < 1. \end{cases} \end{aligned} \tag{A.26}$$

When $p+q \geq 1$, then $Y_1 \leq -\Phi^{-1}(p)$ and $Y_2 \leq -\Phi^{-1}(q)$ together imply $Y_1 + Y_2 \leq 0$. Thus the probability is $\Phi_2(-\Phi^{-1}(p), -\Phi^{-1}(q)) - (1-p)(1-q)$ as claimed. When $p+q < 1$, the claimed result follows as illustrated in Figure A.5.

Finally, we can plug in (A.25) and (A.26) into the expression (A.23) we find that

$$I_1 = \begin{cases} 2\pi \left[(1-q) - \frac{(1-q)^2}{2} - (1-q)(1-p)\right] & \text{if } p < q \text{ and } p+q \geq 1 \\ 2\pi \left[(1-q) - \frac{(1-q)^2}{2} - \frac{1-p^2-q^2}{2}\right] & \text{if } p < q \text{ and } p+q \leq 1 \end{cases}$$

Dividing by $I_2$ now gives our desired result:

$$g(p,q) = \begin{cases} \dfrac{(1-q) - \frac{(1-q)^2}{2} - (1-q)(1-p)}{p(1-q)} = \dfrac{q-(1-2p)}{2p} & \text{if } p < q \text{ and } p+q \geq 1 \\ \dfrac{(1-q) - \frac{(1-q)^2}{2} - \frac{1-p^2-q^2}{2}}{p(1-q)} = \dfrac{p}{2(1-q)} & \text{if } p < q \text{ and } p+q \leq 1. \end{cases}$$

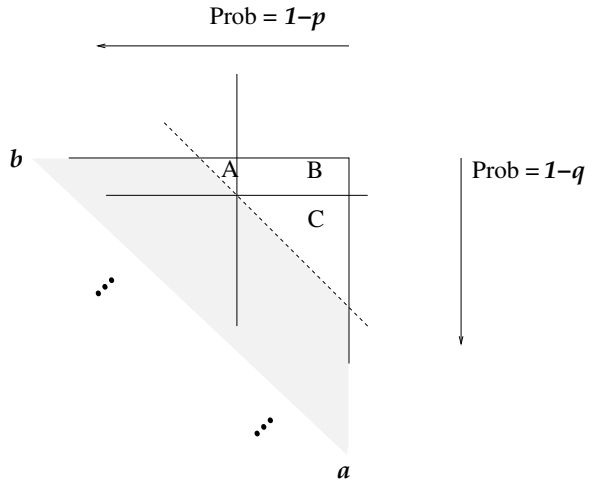



Figure A.5: Area of quadrant $\{Y_1 \leq a, Y_2 \leq b\}$ is $(1-p)(1-q)$. Subtract from this areas $A, B$ and $C$, which are respectively $(1/2-p)^2/2, (1/2-p)(1/2-q)$ and $(1/2-q)^2/2$

## A.6 Supplement for Chapter 7

# Appendix

### A.6.1 Proof of Theorem 7.2.1

i) The law of total expectation gives:

$$\mathbb{E}(\mathcal{X}'') = \mathbb{E}[\mathbb{E}(Y|\mathcal{X}'')] = \mathbb{E}(Y) = \mu_0.$$

ii) Recall that $\mathcal{X}'' = \mathbb{E}(Y|\mathcal{F}'')$, $\mathcal{X}'' \in \mathcal{F}''$, and $\mathcal{F}'' = \sigma(X_1, \ldots, X_N)$. Then,

$$\begin{aligned} &\mathbb{E}(Y|\mathcal{X}'') \\ &= \mathbb{E}[\mathbb{E}(Y|\mathcal{X}'', \mathcal{F}'')|\mathcal{X}''] \qquad \text{(as } \mathcal{X}'' \in \mathcal{F}'') \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}(Y|\mathcal{F}'')|\mathcal{X}''] \\
&= \mathbb{E}(\mathcal{X}''|\mathcal{X}'') \\
&= \mathcal{X}''.
\end{aligned}
$$

iii) This relies on the observation that $\sigma(X_m) = \mathcal{F}_m \subseteq \mathcal{F}'' = \sigma(X_1, \ldots, X_N)$. Then,

$$
\begin{aligned}
\delta_{max} &= \mathrm{Var}\,(X_m) \\
&= \mathbb{E}\left(X_m^2\right) - \mu_0^2 \\
&= \mathbb{E}[\mathbb{E}(Y|\mathcal{F}_m)X_m] - \mu_0^2 && \text{(as } X_m = \mathbb{E}(Y|\mathcal{F}_m)\text{)} \\
&= \mathbb{E}\{\mathbb{E}[\mathbb{E}(Y|\mathcal{F}'')|\mathcal{F}_m]X_m\} - \mu_0^2 && \text{(the smallest } \sigma\text{-field wins)} \\
&= \mathbb{E}[\mathbb{E}(\mathcal{X}''|\mathcal{F}_m)X_m] - \mu_0^2 \\
&= \mathbb{E}[\mathbb{E}(\mathcal{X}''X_m|\mathcal{F}_m)] - \mu_0^2 \\
&= \mathbb{E}(\mathcal{X}''X_m) - \mu_0^2 && \text{(reverse iterated expectation)} \\
&= \mathbb{E}[(\mathcal{X}'' - \mu_0)(X_m - \mu_0)] \\
&\leq \sqrt{\mathrm{Var}\,(\mathcal{X}'')\delta_{max}} && \text{(by the Cauchy-Schwarz inequality).}
\end{aligned}
$$

Squaring and diving both sides by $\delta_{max}$ gives the desired result.

□

## A.6.2 Proof of Theorem 7.2.2

Items ii) and iii) are generalizations of the proof in Ranjan and Gneiting (2010).

i) This follows from direct computation:

$$\mathbb{E}(\mathcal{X}_w) = \mathbb{E}(\mathbf{w}'\mathbf{X}) = \mathbf{w}'\mathbb{E}(\mathbf{X}) = \mu_0\mathbf{w}'\mathbf{1}_N = \mu_0.$$

ii) Consider some reliable aggregate $\mathcal{X}$ such that $\mathbb{E}(Y|\mathcal{X}) = \mathcal{X}$. Then,

$$\begin{aligned}
&\mathbb{E}[(Y-\mathcal{X})^2] \\
&= \mathbb{E}\left\{\mathbb{E}\left[(Y-\mathcal{X})^2|\mathcal{X}\right]\right\} \\
&= \mathbb{E}\left[\mathbb{E}\left(Y^2 - 2Y\mathcal{X} + \mathcal{X}^2|\mathcal{X}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(Y^2|\mathcal{X}\right) - \mathcal{X}^2\right] \\
&= \mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}^2).
\end{aligned}$$

The rest of the proof shows that if $\mathcal{X} = \mathcal{X}_w = \mathbf{w}'\mathbf{X}$, then the above identity cannot hold. This gives a contradiction and hence proves the desired result. First, note that $\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j = 1$. Then,

$$\begin{aligned}
&\mathbb{E}\left[(Y-\mathcal{X}_w)^2\right] \\
&= \mathbb{E}\left[(Y-\mathbf{w}'\mathbf{X})^2\right] \\
&= \mathbb{E}\left\{\left[\sum_{j=1}^{N} w_j(Y-X_j)\right]^2\right\} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}[(Y-X_i)(Y-X_j)] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left(Y^2 - YX_i - YX_j + X_jX_i\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\mathbb{E}\left(Y^2|X_i\right) - \mathbb{E}\left(YX_i|X_i\right) - \mathbb{E}\left(YX_j|X_j\right) + X_jX_i\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\mathbb{E}\left(Y^2|X_i\right) - X_i^2 - X_j^2 + X_jX_i\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\mathbb{E}\left(Y^2|X_i\right) + (X_jX_i - X_jX_i) - X_i^2 - X_j^2 + X_jX_i\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\mathbb{E}\left(Y^2|X_i\right) - X_jX_i - (X_i-X_j)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\mathbb{E}\left(Y^2|X_i\right) - X_j X_i\right] - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\left(X_i - X_j\right)^2\right] \\
&= \mathbb{E}\left(Y^2\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left(X_j X_i\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\left(X_i - X_j\right)^2\right] \\
&= \mathbb{E}\left(Y^2\right) - \mathbb{E}\left(\mathbf{w}'\mathbf{X}\mathbf{X}'\mathbf{w}\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\left(X_i - X_j\right)^2\right] \\
&= \left[\mathbb{E}\left(Y^2\right) - \mathbb{E}\left(\mathcal{X}_w^2\right)\right] - \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[\left(X_i - X_j\right)^2\right].
\end{aligned}
$$

This leads to a contradiction because the double sum on the final line is strictly positive as long as there exists a forecast pair $i \neq j$ such that $\mathbb{P}(X_i \neq X_j) > 0$ and $w_i, w_j > 0$.

iii) The fact that $\mathbb{E}(\mathcal{X}'_w) = \mu_0$ follows similarly to the proof of item i) of Theorem 7.2.1. This item continues under the conditions of the previous item. Therefore it can be assumed that $\mathcal{X}_w$ is not calibrated, that is, $\mathbb{P}(\mathcal{X}'_w \neq \mathcal{X}_w) > 0$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[\left(Y - \mathcal{X}_w\right)^2\right] \\
&= \mathbb{E}\left(Y^2 - 2Y\mathcal{X}_w + \mathcal{X}_w^2\right) \\
&= \mathbb{E}\left(Y^2 + 2\left(\mathcal{X}_w'^2 - \mathcal{X}_w'^2\right) - 2Y\mathcal{X}_w + \mathcal{X}_w^2\right) \\
&= \mathbb{E}\left(Y^2 - 2Y\mathcal{X}'_w + 2\mathcal{X}_w'^2 - 2\mathcal{X}'_w\mathcal{X}_w + \mathcal{X}_w^2\right) \\
&= \mathbb{E}\left[\left(Y - \mathcal{X}'_w\right)^2\right] + \mathbb{E}\left[\left(\mathcal{X}_w - \mathcal{X}'_w\right)^2\right] \\
&= \mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}_w'^2) + \mathbb{E}\left[\left(\mathcal{X}_w - \mathcal{X}'_w\right)^2\right] \qquad \text{(because } \mathcal{X}'_w \text{ is reliable)} \\
&> \mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}_w'^2).
\end{aligned}
$$

Furthermore, from the previous item, $\mathbb{E}\left[\left(Y - \mathcal{X}_w\right)^2\right] < \mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}_w^2)$. Putting this all together gives

$$
\mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}_w'^2) < \mathbb{E}(Y^2) - \mathbb{E}(\mathcal{X}_w^2)
$$

$$\Leftrightarrow \qquad \mathbb{E}(\mathcal{X}_w'^2) - \mu_0^2 > \mathbb{E}(\mathcal{X}_w^2) - \mu_0^2$$

$$\Leftrightarrow \qquad \mathrm{Var}\,(\mathcal{X}_w') > \mathrm{Var}\,(\mathcal{X}_w^2).$$

iv) The fact that $\mathrm{Var}\,(\mathcal{X}_w) \leq \delta_{max}$ follows from direct computation:

$$\begin{aligned}
\mathrm{Var}\,(\mathcal{X}_w) &= \mathbb{E}[(\mu_0 - \mathcal{X}_w)^2] \\
&= \mathbb{E}(\mathcal{X}_w^2) - \mu_0^2 \\
&= \mathbf{w}'\mathbb{E}(\mathbf{X}\mathbf{X}')\mathbf{w} - \mathbf{w}'\mathbf{1}_N\mu_0^2\mathbf{1}_N'\mathbf{w} \\
&= \mathbf{w}'\left[\mathbb{E}(\mathbf{X}\mathbf{X}') - \mu_0^2\mathbf{1}_N\mathbf{1}_N'\right]\mathbf{w} \\
&= \mathbf{w}'\mathbb{E}[(\mathbf{X} - \mathbf{1}_N\mu_0)(\mathbf{X} - \mathbf{1}_N\mu_0)']\mathbf{w} \\
&= \mathbf{w}'\mathrm{Cov}\,(\mathbf{X})\mathbf{w} \\
&\leq \delta_{max}\mathbf{1}_N'\mathbf{w} \\
&= \delta_{max}.
\end{aligned}$$

To see the identity part of the statement, note that

$$\mathrm{Var}\,(\mathcal{X}_w) = \mathbf{w}'\mathrm{Cov}\,(\mathbf{X})\mathbf{w} = \Sigma_{i=1}^N\Sigma_{j=1}^N w_{ij}\mathrm{Cov}\,(X_i, X_j),$$

where $w_{ij} = w_i w_j \in [0,1]$ and $\sum_{i=1}^N \sum_{j=1}^N w_{ij} = 1$. First, suppose that $\mathrm{Var}\,(X_m) = \delta_{max} > \mathrm{Var}\,(X_i) = \delta_i$ for all $i \neq m$. Then, if $w_{ii} > 0$ for some $i \neq m$, the term $w_{ii}\mathrm{Cov}\,(X_i, X_i)$ brings $\mathrm{Var}\,(\mathcal{X}_w)$ below $\delta_{max}$. This decrease cannot be compensated by any other term because no element in $\mathrm{Cov}\,(\mathbf{X})$ is larger than $\delta_{max}$. Consequently, it must be case that $w_i = 0$ for all $i \neq m$. Now, if there exists $j \neq m$ such that $\delta_j = \delta_{max}$ and $w_j > 0$, then $\mathrm{Var}\,(\mathcal{X}_w) = \delta_{max}$ only if all weight is given to $X_m$ and $X_j$, and $\mathrm{Cov}\,(X_j, X_m) = \delta_{max}$. This covariance implies that $\mathrm{Corr}\,(X_j, X_m) = 1$. Thus, $\sigma(X_j) = \sigma(X_m)$ and hence that $X_j = \mathbb{E}[Y|\sigma(X_j)] = \mathbb{E}[Y|\sigma(X_m)] = X_m$.

Consequently, $\mathrm{Var}\left(\mathcal{X}_w\right) = \delta_{max}$ only if all weight is distributed among $X_i$ such that $X_i = X_m$.

From the Theorem 7.2.1, $\delta_{max} \leq \mathrm{Var}\left(\mathcal{X}''\right)$, where the inequality arises from the Cauchy-Schwarz inequality. It is well-known that this reduces to an equality if and only if $\mathcal{X}''$ and $X_m$ are linearly dependent. Such a linear dependence would imply that $\sigma(\mathcal{X}'') = \sigma(X_m)$ and hence that $X_m = \mathbb{E}[Y|\sigma(X_m)] = \mathbb{E}[Y|\sigma(\mathcal{X}'')] = \mathcal{X}''$. Now, if there exists $j \neq m$ such that $\delta_j = \delta_{max}$, then by the same argument $\sigma(\mathcal{X}'') = \sigma(X_m) = \sigma(X_j)$ and consequently $X_j = X_m = \mathcal{X}''$.

Putting this all together gives that $\mathbf{w}'\mathbf{X} = \mathcal{X}''$ if and only if $\sigma(X_m) = \sigma(\mathcal{X}'')$ and $w_i > 0$ only for all $X_i = X_m$.

□

### A.6.3 Derivation of Equation 7.4

Suppose that $\mathcal{X}_k \in \{f_1, \ldots, f_I\}$ for some finite $I$. Let $K_i$ be the number of times $f_i$ occurs, $\bar{Y}_i$ be the empirical average of $\{Y_k : \mathcal{X}_k = f_i\}$, and $\bar{Y} = \frac{1}{K}\sum_{k=1}^{K} Y_k$. Then,

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}(Y_k - \mathcal{X}_k)^2 \\
&= \frac{1}{K}\left(\sum_{k=1}^{K}\mathcal{X}_k^2 - 2\sum_{k=1}^{K}Y_k\mathcal{X}_k + \sum_{k=1}^{K}Y_k^2\right) \\
&= \frac{1}{K}\left[\sum_{i=1}^{I}K_i f_i^2 - 2\sum_{i=1}^{I}K_i f_i\bar{Y}_i + \left(2\sum_{i=1}^{I}K_i\bar{Y}_i\bar{Y} - 2\sum_{i=1}^{I}K_i\bar{Y}_i\bar{Y}\right)\right. \\
&\qquad \left. + \left(\sum_{i=1}^{I}K_i\bar{Y}^2 - \sum_{i=1}^{I}K_i\bar{Y}^2\right) + \sum_{k=1}^{K}Y_k^2\right] \\
&= \frac{1}{K}\left[\sum_{i=1}^{I}K_i\left(f_i^2 - 2f_i\bar{Y}_i + 2\bar{Y}_i\bar{Y} - \bar{Y}^2\right) + \sum_{k=1}^{K}(Y_k^2 - 2\bar{Y}_k\bar{Y} + \bar{Y}^2)\right] \\
&= \frac{1}{K}\left[\sum_{i=1}^{I}K_i\left(f_i^2 - 2f_i\bar{Y}_i + (\bar{Y}_i^2 - \bar{Y}_i^2) + 2\bar{Y}_i\bar{Y} - \bar{Y}^2\right) + \sum_{k=1}^{K}(Y_k - \bar{Y})^2\right]
\end{aligned}
$$

$$= \frac{1}{K}\left[\sum_{i=1}^{I} K_i\left(f_i^2 - 2f_i\bar{Y}_i + \bar{Y}_i^2\right) - \sum_{i=1}^{I} K_i\left(\bar{Y}_i^2 - 2\bar{Y}_i\bar{Y} + \bar{Y}^2\right) + \sum_{k=1}^{K}(Y_k - \bar{Y})^2\right]$$

$$= \frac{1}{K}\sum_{i=1}^{I} K_i\left(f_i - \bar{Y}_i\right)^2 - \frac{1}{K}\sum_{i=1}^{I} K_i\left(\bar{Y}_i - \bar{Y}\right)^2 + \frac{1}{K}\sum_{k=1}^{K}(Y_k - \bar{Y})^2.$$

# Bibliography


Allard, D., Comunian, A., and Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, 44:545–581.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., and Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6(2):130.

Armstrong, J. S. (2001). Combining forecasts. In Armstrong, J. S., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 417–439. Kluwer Academic Publishers, Norwell, MA.

Ashton, A. H. and Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31(12):1499–1508.

Atanasov, P. D., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P. E., Ungar, L., and Mellers, B. (2015). Distilling the wisdom of crowds: Prediction markets versus prediction polls. *Management Science, Forthcoming*.

Baars, J. A. and Mass, C. F. (2005). Performance of national weather service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and Forecasting*, 20(6):1034–1047.

Bacharach, M. (1972). Scientific disagreement. Unpublished manuscript.

Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *Information Theory, IEEE Transactions on*, 51(7):2664–2669.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.

Batchelder, W. H., Strashny, A., and Romney, A. K. (2010). Cultural consensus theory: Aggregating continuous responses in a finite interval. In *Advances in Social Computing*, pages 98–107. Springer.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Or*, pages 451–468.

Bier, V. (2004). Implications of the research on expert overconfidence and dependence. *Reliability Engineering & System Safety*, 85(1):321–329.

Blix, M., Wadefjord, J., Wienecke, U., and Adahl, M. (2001). How good is the forecasting performance of major institutions? *sveriges riksbank economic review*, pages 38–68.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilitá. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, 10:1137–1148.

Braun, P. A. and Yaniv, I. (1992). A case study of expert judgment: Economists' probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making*, 5(3):217–231.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bröcker, J. and Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661.

Broomell, S. B. and Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74(3):531–553.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Manuscript available at www-stat.wharton.upenn.edu/ buja.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Cedilnik, A., Kosmelj, K., and Blejec, A. (2004). The distribution of the ratio of jointly normal variables. *Metodoloski Zvezki*, 1(1):99–108.

Chen, Y. (2009). Learning classifiers from imbalanced, only positive and unlabeled data sets. *Department of Computer Science Iowa State University*.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–583.

Clemen, R. T. and Winkler, R. L. (2007). Aggregating probability distributions. *Advances in Decision Analysis*, pages 154–176.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York, NY, USA.

Dawid, A., DeGroot, M., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M., Lindley, D., McConway, K., and Winkler, R. (1995). Coherent combination of experts' opinions. *TEST*, 4(2):263–313.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

DeGroot, M. H. (1988). A bayesian view of assessing uncertainty and comparing expert opinion. *Journal of statistical planning and inference*, 20(3):295–306.

DeGroot, M. H. and Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, 37(5):546–558.

DeSart, J. (2012 (accessed November 3, 2015)). *Presidential Election Forecasting*. `http://research.uvu.edu/DeSart/forecasting`.

Di Bacco, M., Frederic, P., and Lad, F. (2003). Learning from the probability assertions of experts. Research Report. Available at: http://www.math.canterbury.ac.nz/research/ucdms2003n6.pdf.

Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38.

Dowie, J. (1976). On the efficiency and equity of betting markets. *Economica*, 43(170):139–150.

Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

Elliott, G. and Timmermann, A. (2013). *Handbook of Economic Forecasting SET 2A-2B*. Elsevier.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527.

Ernst, P., Pemantle, R., Satopää, V. A., and Ungar, L. H. (2016). Bayesian aggregation of two forecasts in the partial information framework. *Statistics & Probability Letters (Under Review)*.

Flores, B. E. and White, E. M. (1989). Subjective versus objective combining of forecasts: an experiment. *Journal of Forecasting*, 8(3):331–341.

Forlines, C., Miller, S., Prakash, S., and Irvine, J. (2012). Heuristics for improving forecast aggregation. In *AAAI Fall Symposium: Machine Aggregation of Human Judgment*.

Foster, D. P. and Vohra, R. V. (1998). Asymptotic calibration. *Biometrika*, 85(2):379–390.

Fox, C. R. and Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14:195–200.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. CRC press.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Institute of Electrical and Electronics Engineer (IEEE) Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148.

Gent, I. P. and Walsh, T. (1996). Phase transitions and annealed theories: Number partitioning as a case study. In *Proceedings of European Conference on Artificial Intelligence (ECAI 1996)*, pages 170–174. John Wiley & Sons.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, 98(4):506.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Rejoinder on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):256–264.

Goel, S., Reeves, D. M., Watts, D. J., and Pennock, D. M. (2010). Prediction without markets. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 357–366. ACM.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 14(1):107–114.

Graefe, A., Armstrong, J. S., Jones, R. J., and Cuzán, A. G. (2014a). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54.

Graefe, A., Küchenhoffb, H., Stierleb, V., and Riedlb, B. (2014b). Combining forecasts: Evidence on the relative accuracy of the simple average and bayesian model averaging for predicting social science problems.

Gubin, L., Polyak, B., and Raik, E. (1967). The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24.

Hastings, C., Mosteller, F., Tukey, J. W., and Winsor, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426.

Hayes, B. (2002). The easiest hard problem. *American Scientist*, 90(2):113–117.

Hibon, M. and Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*, 21(1):15–24.

Hong, L. and Page, S. (2009). Interpreted and generated signals. *Journal of Economic Theory*, 144(5):2174–2196.

Hwang, J. and Pemantle, R. (1997). Estimating the truth of an indicator function of a statistical hypothesis under a class of proper loss functions. *Statistics & Decisions*, 15:103–128.

Hwang, S.-G. (2004). Cauchy's interlace theorem for eigenvalues of hermitian matrices. *American Mathematical Monthly*, pages 157–159.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, 5(1):55–71.

Karmarkar, N. and Karp, R. M. (1982). *The Differencing Method of Set Partitioning*. Computer Science Division, University of California Berkeley.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, 21(1):61–72.

Kellerer, H., Pferschy, U., and Pisinger, D. (2004). *Knapsack Problems*. Springer.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, 39(1):98–114.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution.

Lai, T. L., Gross, S. T., Shen, D. B., et al. (2011). Evaluating probability forecasts. *The Annals of Statistics*, 39(5):2356–2382.

Langford, E., Schwertman, N., and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325.

Laurent, M., Poljak, S., and Rendl, F. (1997). Connections between semidefinite relaxations of the max-cut and stable set problems. *Mathematical Programming*, 77(1):225–246.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54(4):421–431.

Lichman, M. (2013). UCI machine learning repository.

Lichtendahl Jr, K. C. and Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755.

Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance*, 20(2):159–183.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). *Calibration of probabilities: The state of the art*. Springer.

Lobo, M. S. and Yao, D. (2010). Human judgement is heavy tailed: Empirical evidence and implications for the aggregation of estimates and forecasts. *Fontainebleau: INSEAD*.

Lubinski, D. and Humphreys, L. G. (1996). Seeing the forest from the trees: When predicting the behavior or status of groups, correlate means. *Psychology, Public Policy, and Law*, 2(2):363.

McCullagh, P., Nelder, J. A., and McCullagh, P. (1989). *Generalized linear models*, volume 2. Chapman and Hall London.

McKenzie, C. R., Liersch, M. J., and Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, 107(2):179 – 191.

McMullen, P. and Shephard, G. C. (1971). *Convex Polytopes and the Upper Bound Conjecture*, volume 3. Cambridge University Press, Cambridge, U.K.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:10871092.

Meyer, C. (2009). The bivariate normal copula. *Communications in Statistics - Theory and Methods*, 42:2402–2422.

Migon, H. S., Gamerman, D., Lopes, H. F., and Ferreira, M. A. (2005). Dynamic models. *Handbook of Statistics*, 25:553–588.

Mills, T. C. (1991). *Time Series Techniques for Economists*. Cambridge University Press.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

Moore, D. A. and Klein, W. M. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes*, 107(1):60–74.

Morgan, M. G. (1992). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press.

Murphy, A. and Winkler, R. (1987a). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600.

Murphy, A. H. and Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts. comparison of results from the first and second years of the zierikzee experiment. *Monthly Weather Review*, 112(3):413–423.

Murphy, A. H. and Winkler, R. L. (1977a). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature. *National Weather Digest*, 2(2):2–9.

Murphy, A. H. and Winkler, R. L. (1977b). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1):41–47.

Murphy, A. H. and Winkler, R. L. (1987b). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–741.

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)*. Princeton University Press.

Pal, S. (2009). A note on a conjectured sharpness principle for probabilistic forecasting with calibration. *Biometrika*, pages 1019–1023.

Papadatos, N. (1995). Maximum variance of order statistics. *Annals of the Institute of Statistical Mathematics*, 47(1):185–193.

Parunak, H. V. D., Brueckner, S. A., Hong, L., Page, S. E., and Rohwer, R. (2013). Characterizing and aggregating agent estimates. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, pages 1021–1028, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford, U.K.

Plackett, R. (1954). A reduction formula for normal multivariate integrals. *Biometrika*, 41:351–360.

Polyakova, E. I. and Journel, A. G. (2007). The nu expression for probabilistic data integration. *Mathematical Geology*, 39:715–733.

Primo, C., Ferro, C. A., Jolliffe, I. T., and Stephenson, D. B. (2009). Calibration of probabilistic forecasts of binary events. *Monthly Weather Review*, 137(3):1142–1149.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Ranjan, R. (2009). *Combining and Evaluating Probabilistic Forecasts*. PhD thesis, University of Washington.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.

Rao, C. R. (2009). *Linear Statistical Inference and Its Applications*, volume 22 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, New York.

Ravishanker, N. and Dey, D. K. (2001). *A first course in linear model theory*. CRC Press.

Rowse, G. L., Gustafson, D. H., and Ludke, R. L. (1974). Comparison of rules for aggregating subjective likelihood ratios. *Organizational Behavior and Human Performance*, 12(2):274–285.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, 2(2):191–201.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Supplement to "probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs".

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., Ungar, L. H., et al. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, 8(2):1256–1280.

Satopää, V. A., Jensen, S. T., Pemantle, R., and Ungar, L. H. (2016). Partial information framework: Aggregating estimates from diverse information sources. *Journal of the American Statistical Association (arXiv:1505.06472) (Under Review)*.

Satopää, V. A., Pemantle, R., and Ungar, L. H. (2015). Modeling probability forecasts via information diversity. *The Journal of the American Statistical Association (Theory & Methods) (arXiv:1406.2148) (In Press)*.

Satopää, V. A. and Ungar, L. H. (2015). Combining and extremizing real-valued forecasts. *arXiv:1506.06405 (Under Review)*.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Shlomi, Y. and Wallsten, T. S. (2010). Subjective recalibration of advisors' probability estimates. *Psychonomic Bulletin & Review*, 17(4):492–498.

Shlyakhter, A. I., Kammen, D. M., Broido, C. L., and Wilson, R. (1994). Quantifying the credibility of energy projections from trends in past data: The US energy sector. *Energy Policy*, 22(2):119–130.

Silver, N. (2012 (accessed November 3, 2015)). *FiveThirtyEight's 2012 Forecast*. http://fivethirtyeight.blogs.nytimes.com/fivethirtyeights-2012-forecast.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65(2):117–137.

Tanaka, M. and Nakata, K. (2014). Positive definite matrix approximation with condition number constraint. *Optimization Letters*, 8(3):939–947.

Tanner Jr, W. P. and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6):401.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, New Jersey.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. The Association for the Advancement of Artificial Intelligence Technical Report FS-12-06.

Vislocky, R. L. and Fritsch, J. M. (1995). Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society*, 76(7):1157–1164.

Wallace, B. C. and Dahabreh, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *Institute of Electrical and Electronics Engineers (IEEE) 12th International Conference on Data Mining (International Conference on Data Mining)*, pages 695–704. Institute of Electrical and Electronics Engineers (IEEE).

Wallsten, T. S., Budescu, D. V., and Erev, I. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10:243–268.

Wallsten, T. S. and Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, 41(1):1–18.

Wilson, A. G. (1994). *Cognitive Factors Affecting Subjective Probability Assessment*. Citeseer.

Wilson, P. W., DAgostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847.

Winkler, R. L. and Jose, V. R. R. (2008). Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):251–255.

Winkler, R. L. and Murphy, A. H. (1968). Good probability assessors. *Journal of Applied Meteorology*, 7(5):751–758.

Won, J. H. and Kim, S.-J. (2006). Maximum likelihood covariance estimation with a condition number constraint. In *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pages 1445–1449. IEEE.

Wright, G., Rowe, G., Bolger, F., and Gammack, J. (1994). Coherence, calibration, and expertise in judgmental probability forecasting. *Organizational Behavior and Human Decision Processes*, 57(1):1–25.

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Inc.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808.

Zhang, H. and Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, 6:1–14.

Ziegler, G. M. (2000). Lectures on 0/1-polytopes. In Kalai, G. and Ziegler, G. M., editors, *Polytopes - Combinatorics and Computation*, volume 29, pages 1–41, Basel. Springer, Birkhäuser.